ASSESSMENT AND COMPARISON OF TECHNIQUES FOR TRANSFORMING PARAMETERS TO A COMMON METRIC IN ITEM RESPONSE THEORY

BY

DANIEL OWEN SEGALL

B.A., University of California, 1979
A.M., University of Illinois, 1981

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1983

Urbana, Illinois

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release: distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Measurement Series 84-6	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Michael V. Levine Model-Based Measurement Lab 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Cognitive Science Research Program Office of Naval Research
6c. ADDRESS (City, State, and ZIP Code) University of Illinois 210 Education Building, 1310 S. Sixth St. Champaign, IL 61820	7b. ADDRESS (City, State, and ZIP Code) Code 1142CS 800 North Quincy St. Arlington, VA 22217
8a. NAME OF FUNDING / SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO: 61153N PROJECT NO.: RR042-04 TASK NO: RR04204-01 WORK UNIT ACCESSION NO.: NR4421-546 NR 150-518
11. TITLE (Include Security Classification) Assessment and Comparison of Techniques for Transforming Parameters to a Common Metric in Item Response Theory	
12. PERSONAL AUTHOR(S) Daniel Owen Segall	
13a. TYPE OF REPORT 13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (Year, Month, Day) 1988 15. PAGE COUNT
16 SUPPLEMENTARY NOTATION	

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Item response theory, linking, equating, scale transformation, common metric.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Seven different techniques for transforming two sets of parameters to a common metric in item response theory were evaluated. These techniques include three techniques that estimate the transformation by equating the first two moments of the distributions of estimated difficulty parameters. Another three techniques are included that estimate the transformation by minimizing sums of squares criteria. The remaining technique (MLE) finds the scale transformation that maximizes the likelihood of observing vectors of item parameter differences. Results indicate that for calibrations with a small number of items with extreme estimated difficulty values all techniques perform satisfactorily for moderate sample sizes (500 subjects) and test lengths (60 items). With a large number of extreme estimated difficulty values the three sums of squares and MLE techniques perform satisfactorily, while the difficulty parameter techniques do not. The MLE technique appears poorly suited for smaller samples (500 subjects) and shorter tests (30 items).

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION
22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code) / 22c OFFICE SYMBOL

ASSESSMENT AND COMPARISON OF TECHNIQUES FOR TRANSFORMING PARAMETERS TO A COMMON METRIC IN ITEM RESPONSE THEORY

Daniel Owen Segall, Ph.D.
Department of Psychology
University of Illinois at Urbana-Champaign, 1983

Seven different techniques for transforming two sets of parameters to a common metric in item response theory were evaluated. These techniques include three techniques that estimate the transformation by equating the first two moments of the distributions of estimated difficulty parameters. Another three techniques are included that estimate the transformation by minimizing sums of squares criteria. The remaining technique (MLE) finds the scale transformation that maximizes the likelihood of observing vectors of item parameter differences. Results indicate that for calibrations with a small number of items with extreme estimated difficulty values all techniques perform satisfactorily for moderate sample sizes (500 subjects) and test lengths (60 items). With a large number of extreme estimated difficulty values the three sums of squares and MLE techniques perform satisfactorily, while the difficulty parameter techniques do not. The MLE technique appears poorly suited for smaller samples (500 subjects) and shorter tests (30 items).

ACKNOWLEDGEMENTS

I would like to thank Drs. Fritz Drasgow, Charles Hulin, Lloyd Humphreys, Michael Levine and Robert Linn for their time and effort as members of my thesis committee.

There are three individuals who have made special contributions to my graduate education. Dr. Charles Hulin has provided special support and guidance throughout my graduate training. He has been a great inspiration, academically and professionally. Dr. Fritz Drasgow has provided much professional guidance and has been a tremendous resource. He has devoted large amounts of time and effort on numerous projects, and has offered a great deal of encouragement. Dr. Michael Levine has provided me with a unique opportunity for intellectual growth and stimulation. He has invested countless hours in personal consultation, and has been a primary force in shaping my intellectual and professional development.

ASSESSMENT AND COMPARISON OF TECHNIQUES FOR TRANSFORMING PARAMETERS TO A COMMON METRIC IN ITEM RESPONSE THEORY

BY

DANIEL OWEN SEGALL

B.A., University of California, 1979
A.M., University of Illinois, 1981

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1983

Urbana, Illinois

TABLE OF CONTENTS

CHAPTER PAGE

CHAPTER 1

INTRODUCTION

Item Response Theory (IRT) enjoys several theoretical advantages over earlier theories of psychological measurement. One such characteristic, the property of "item parameter invariance," is the central focus of this study. As Lord (1980, p.35) states "The invariance of item parameters across groups is one of the most important characteristics of item response theory."

The property of "invariance" refers to the parameters of the item response function. The shape of the response function is usually described by either the logistic or normal ogive models (see Table 1). We can see from Table 1 that for any given item with parameters a, b, and c, the relation between the ability parameter (θ) and the probability of a correct answer is fully specified. The probability of a correct answer to a particular item, among examinees selected at random with a given ability level θ_o, depends only on θ_o, not on the number of people at θ_o, or on the number of people at other ability levels. Even if two groups differ in their distributions of ability, their response functions will be the same. Examinees with θ_o from one group have the same probability of passing the item as do examinees with θ_o from the other group. Thus, the probability of a correct response among individuals with ability θ is independent of group membership, and depends only on θ and the item parameters (a, b, and c). The lower asymptote, the point of inflection, and the slope all remain invariant across groups.

Table 1

Three parameter logistic:

$$P = C + \frac{1 - C}{1 + e^{Da(\theta - b)}}$$

Two parameter logistic:

$$P = \frac{1}{1 + e^{Da(\theta - b)}}$$

Three parameter normal ogive:

$$P = C + (1 - C)\int_{-\infty}^{a(\theta - b)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

where D = -1.7

The property of item parameter invariance has led to several important applications of item response theory. One such application has been in the area of item bias. An item is said to be biased when examinees with the same level of the trait from different subpopulations have different probabilities of answering the item correctly. The basic paradigm used in this problem area involves two independent estimates of item parameters, one from each of two groups. Because of the invariance property, any differences in corresponding item parameters between the two groups (beyond those expected by sampling error) would be an indication of bias.

Item parameter invariance has also suggested a powerful technique for the assessment of the quality of a translated item from one language to a new language. The paradigm is identical to that described under item bias, except here, if the response function is found to be group dependent, the quality of the translation is suspected.

A third important application of item response theory, made possible in part by item parameter invariance, is the development of large, pre-calibrated item pools. One of the most appealing procedures for developing these pools involves administering a number of overlapping tests to separate groups of individuals (McKinley & Reckase, 1981). These tests are overlapping in the sense that they have some items in common. These common items provide the link necessary to place all items on the same scale. The development of large pools of items is especially important for tailored testing.

There is one problem, however, that hampers the application of IRT as described above and which provides the impetus for this paper. The

parameters for the logistic and normal ogive models are invariant only up to a linear transformation of the scale of ability. This problem is caused by the indeterminacy of the origin and unit of measure of the ability scale. As Lord explains:

> If a parameter value is in principle indeterminate even when we are given the entire population of observable values, then the parameter is called unidentifiable. Actually, all θ, a_i and b_i (but not c_i) are unidentifiable until we agree on some arbitrary choice of origin and unit of measurement. Once this choice is made, all θ and item parameters will ordinarily be identifiable in a suitable infinite population of examinees and infinite pool of test items. (Lord 1980, p.184-185)

Thus the origin and unit of measure of our ability scale is arbitrary. This causes difficulty when we wish to compare two sets of independently estimated parameters, as outlined in the examples above. Because the decision is arbitrary for each group, there is no assurance that the origin and unit were selected in such a way as to make the two sets of parameters comparable. The purpose of this paper is to investigate techniques that transform one test's metric to the metric of another test and thus permit the direct comparison of all item response functions between the two groups. In addition a new technique for transforming parameters to a common metric is introduced.

Transforming to a Common Metric: Symptoms and Formalization

In the remainder of this chapter we will work through a hypothetical

example that will demonstrate some common symptoms of the equating problem. In the final section of this chapter, the problem will be formalized and the basic transformation equations presented.

Let us suppose in this example that we are dealing with a test of verbal skills and that our two hypothetical populations are all 5th grade and 6th grade students respectively. Let us further suppose that we are interested in examining a vocabulary test for bias between our 5th and 6th grade populations. That is, we are interested in identifying items that function differently for the two grade levels.

Our first step as indicated by Figure 1, would be to select two samples of 5th and 6th grade students. Next, we administer the test to each group to obtain our item responses. From these item responses we obtain independent estimates of item and person parameters, as indicated at the bottom of Figure 1. Let us suppose we used LOGIST (Wood, Wingersky, & Lord, 1976) to obtain our parameter estimates.

The next step, before comparing any parameters directly, would be to transform one set of parameters to the scale of the other set. That is to transform the parameters to a common metric. For now, however, let us observe the consequences of ignoring the equating phase altogether. That is, let us observe symptoms of the equating problem when we attempt to compare parameters directly after estimation.

Figure 2 displays the two hypothetical histograms for our 5th and 6th grade samples that we would expect to obtain from our LOGIST estimated thetas. The ordinate represents the proportion of examinees observed at each level of theta, on a basis of the LOGIST estimated thetas. (In reality we would not expect our observed histograms to be

Figure 1

Outline of Procedure of Hypothetical Example

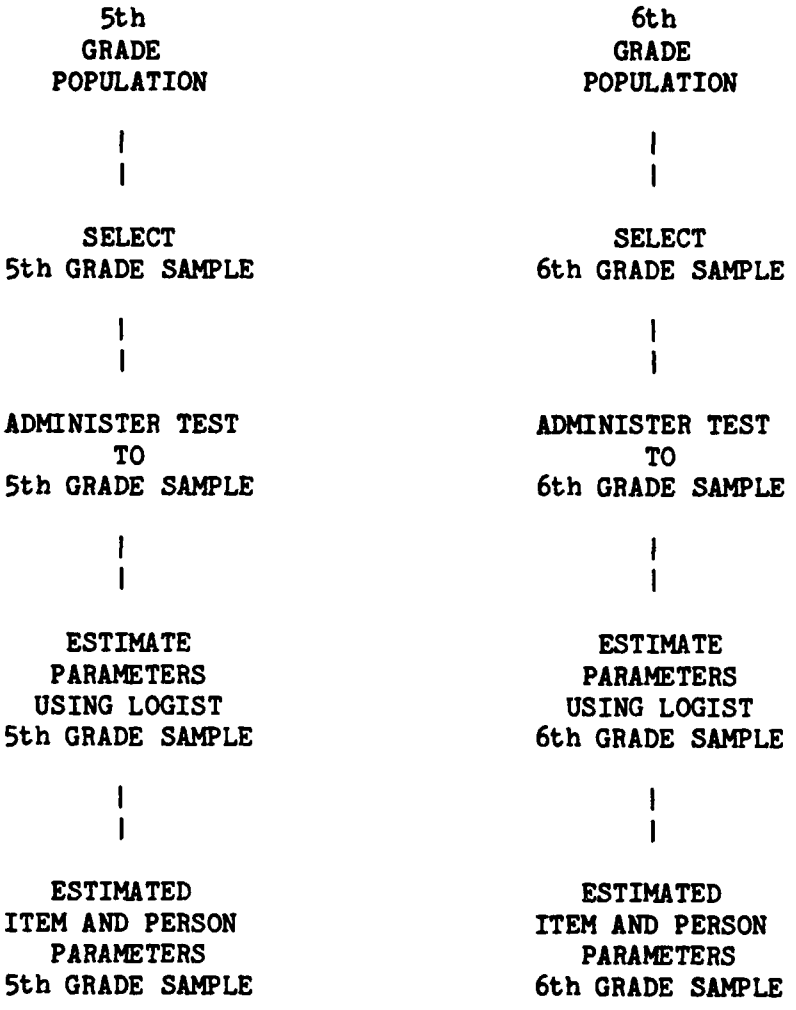
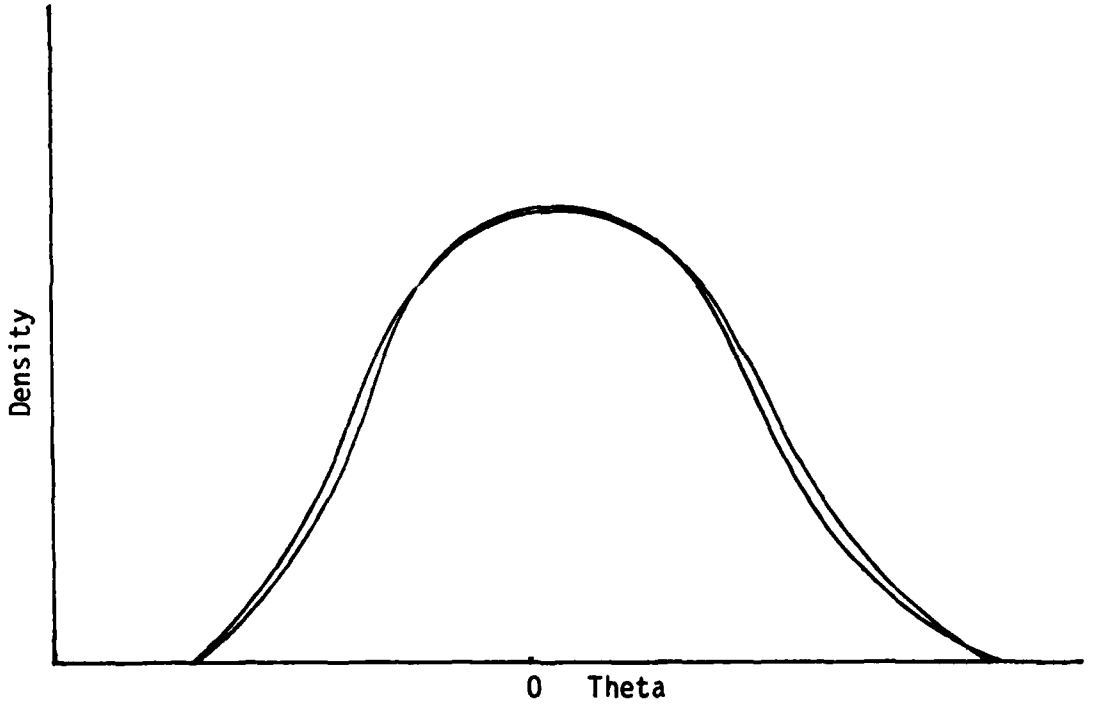


Figure 2

Histograms for 5th and 6th Grade
Ability Estimates from LOGIST
(Example)



as smooth as those in Figure 2) Notice that the two histograms for our two sets of independently estimated ability parameters overlap to a high degree. In fact, just by observing these two distributions, one might be easily convinced that their mean and standard deviations are identical. But this observation does not support what we know about the verbal ability of 5th versus 6th grade students. We would expect, with a great deal of certainty, that the mean of our estimated thetas for the 6th grade sample would be significantly higher than the mean of our 5th grade sample. But when we observe our independent estimates of the ability parameters for our two groups, their mean and standard deviation appear identical.

This latter occurrence is no accident. Remember, that because the origin and unit of measure are arbitrary in item response theory, our estimation procedure is free to select any values. LOGIST selects the origin and unit of the measurement scale such that they correspond to the mean and standard deviation (respectively) of the estimated person parameters. Thus, for each group of independent parameter estimates, the origin of our scale was set to the mean of our estimated thetas, and the unit was set equal to the standard deviation. It is not surprising that the first two moments of our estimated 5th and 6th grade distributions look identical. The scale of ability was selected using these moments as criteria. When these criteria are used to select the origin and unit, no matter to what extent the mean and standard deviation differ between two groups, they will always appear identical.

Now, let us observe the symptoms of the equating problem when we attempt to compare estimated <u>item</u> parameters, without first transforming

to a common metric.

Suppose for the first item of our vocabulary test we obtained the two items with parameter estimates given in Table 2. ICC's for these two curves are given in Figure 3. The two overlapping ability distributions are represented by the u-shaped curves along the base line. By examination of either Table 2 or Figure 3, we might conclude (if we were unaware of an equating problem), that the first item is biased against the 5th grade sample. That is, for most levels of ability, the probability of obtaining a correct response is larger for 6th graders than for the 5th graders. But, when we examine the item parameter estimates for the rest of the items in the test, we notice a similar pattern. The b_i's for the 5th grade sample appear consistantly higher than the 6th grade estimates, while the a_i's are slightly smaller for the 5th grade group. Could every item on the test be biased? Not likely. These problems are all symptoms of side-stepping the equating stage before attempting to compare the two independent sets of parameters.

Let us now consider how to resolve the discrepancy between the two independent estimates of these item parameters through equating to a common metric. Remember that, in accordance with item response theory, the origin and unit of measure for each of our scales is arbitrary. Thus we can apply a linear transformation to either or both of our scales (base lines) and not violate any assumptions of IRT. A convenient terminology has been developed (Linn, Levine, Hastings & Wardrop, 1981) that helps clarify the basic issues in developing a common metric. This terminology involves the arbitrary designation of

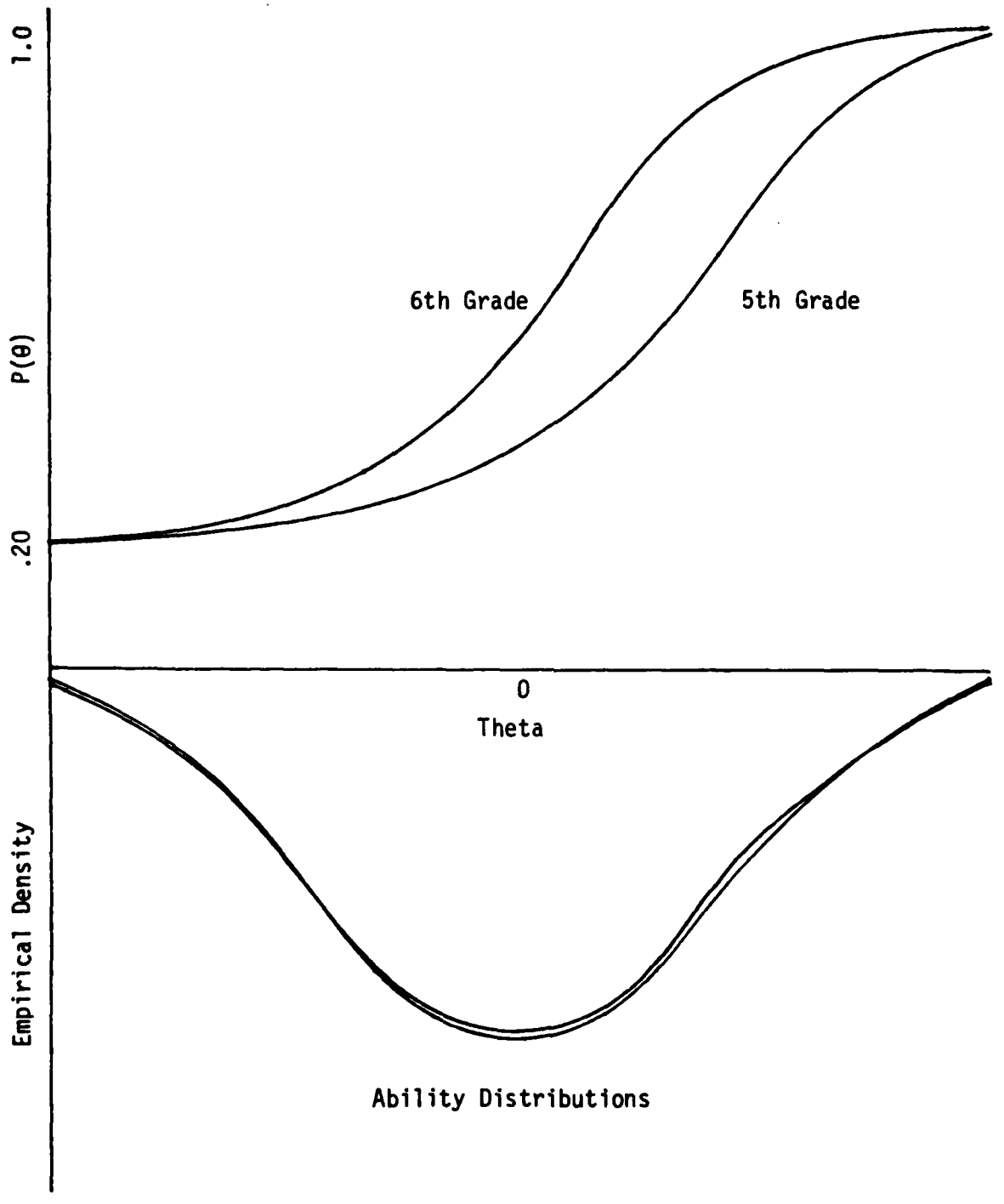
Table 2

Estimated Item Parameters for Hypothetical Example

FIFTH GRADE ITEM PARAMETERS		SIXTH GRADE ITEM PARAMETERS	
a_i	.95	a_i	1.00
b_i	.79	b_i	0.00
c_i	.20	c_i	.20

Figure 3

Item Parameter (first item) and Ability Distribution Estimates from LOGIST (before equating)



one of our two groups as the "base group" and the other as the "comparison group." The scale and parameters of our base group are held fixed, while the parameters of our comparison group are transformed to the scale of the base group. Thus, after transformation of the comparison group parameters, the scale of our base group will be the scale on which all our estimated parameters will be measured.

For now, let us hold the scale of our 6th grade group fixed (thus designating it as the base group) and apply a linear transformation only to the scale of our 5th grade sample (which now becomes our comparison group). Our new scale will be defined as:

$$\theta^* = A \times \theta + B \qquad [1.1]$$

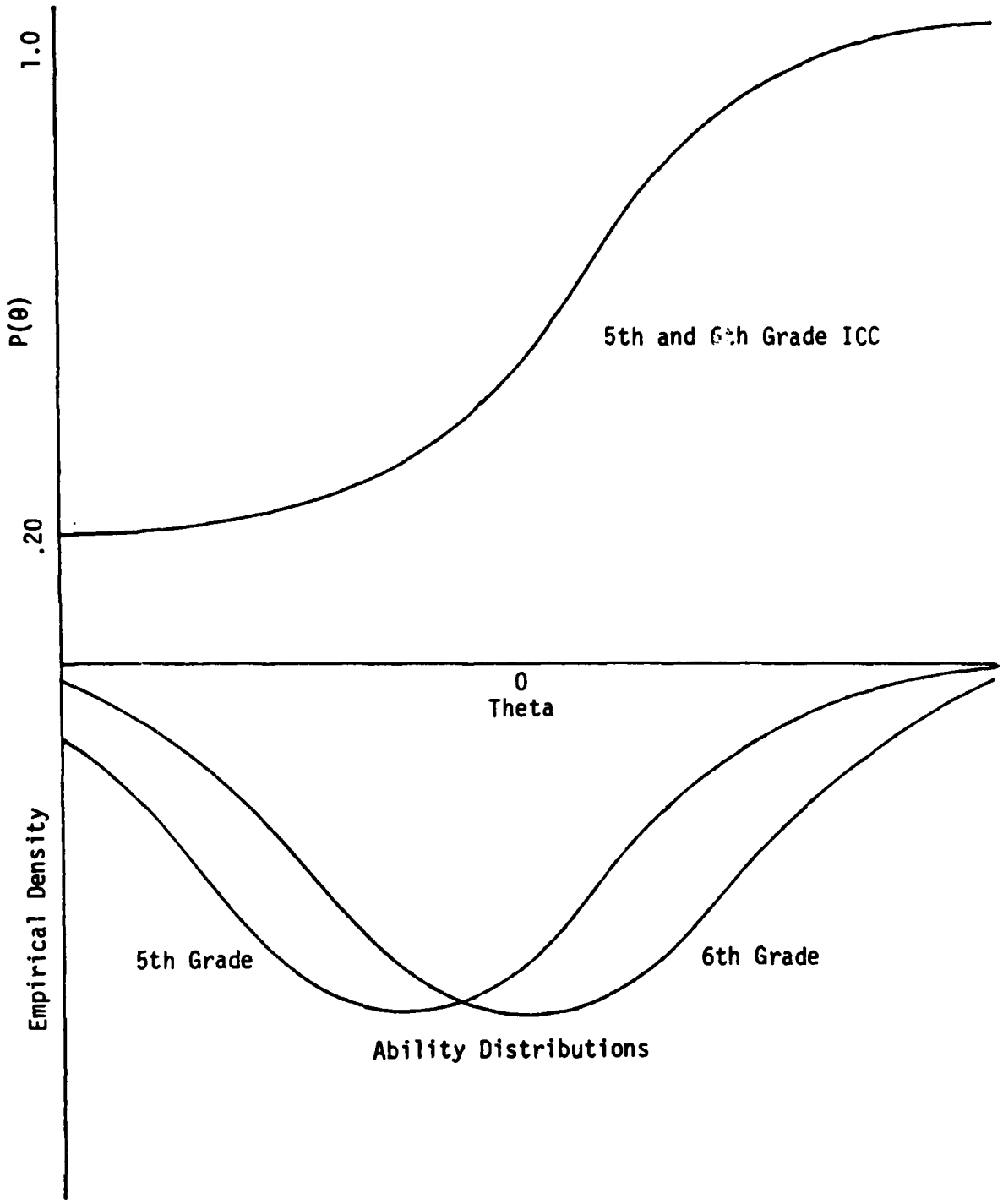
where A and B are the components of the linear transformation that transform points on the old comparison group scale (θ) to points on the new equated base group scale (θ^*).

Returning to the example in Figure 3, we can see that if we move the scale for our comparison group (5th graders) to the left, and then contract it slightly, the two item characteristic curves will line up exactly, as displayed in Figure 4. Notice that as we moved (transformed) the scale for the 5th grade ICC, the ability distribution for that group moved along with it. Thus after the transformation we observe an item characteristic curve that is identical for both groups. We also observe two distinct ability distributions, with our 5th grade sample scoring, on the average, lower than our 6th grade sample, as we would have expected on a basis of prior knowledge.

If this same exact transformation of scale were applied to all the other items of our comparison group, we would observe similar results.

Figure 4

Item Parameter (first item) and Ability Distribution Estimates from LOGIST (after equating)



That is, our ICC's should match up, just as they did in the above example.

As implied by equation [1.1] we can quantify the transformation required to convert the scale of our comparison group parameters (θ) to the metric (θ^*) of our arbitrarily designated base group. In the example above, we found the transformation for our 5th grade sample that placed it on the same metric as our 6th grade sample. Finding the "correct" transformation involves finding the correct values of A and B in equation [1.1]. These particular values will depend on several factors. First, the values of A and B will depend on the rule that our estimation procedure uses for assigning values to the origin and unit of the scale for each set of parameter estimates. If this rule is consistant across data sets, then indirectly this transformation will also depend on the differences in the distributions of ability of the two groups from which our independent estimates were calculated.

Once we have found the correct values for A and B, we can apply the same scale transformation to every item in the test for the group requiring the transformation (comparison group). The point to be stressed here, is that the scale transformation is identical for every item of the test. That is, the values of A and B are not item specific. Rather, they (and the transformation they represent) are constant across items. Because the values of A and B are constant across items, they are often refered to as "equating constants".

After the values of our equating constants (A and B) have been identified, the values of all the item parameters for the comparison group may be transformed to the new scale using the following equations:

$$\underline{a}i^* = \underline{a}i \;/\; A \qquad [1.2a]$$

$$\underline{b}i^* = A \times \underline{b}i + B \qquad [1.2b]$$

$$\underline{c}i^* = \underline{c}i \qquad [1.2c]$$

$$\underline{\theta}a^* = A \times \underline{\theta}a + B \qquad [1.2d]$$

These equations represent the new values of the parameters for the comparison group that place them on a common metric with the base group. It is important to remember that under this paradigm, we are transforming the parameters for only one of our two groups (the comparison group). The parameters for the base group remain completely unaffected by these transformations. That is, their values go unaltered throughout the entire equating process. The asterisk in equations [1.2a-d] represent the values of the comparison group parameters transformed to the base group scale. The subscript i refers to item i; the A and B are the equating constants introduced in equation [1.1]; and the parameters without the asterisk are the comparison group parameters before transforming to the base group metric.

At this point, a few comments about equations [1.2a-d] would be in order. Let us begin with the transformation of $\underline{c}i$. Why does this parameter remain unaffected by a linear transformation of the theta scale? If we refer back to Table 1, we see that for both the three parameter logistic and the normal ogive models, the $\underline{c}i$ represents the probability of making a correct response to item i by a randomly sampled individual with a theta of minus infinity. It is readily apparent that any linear transformation of the theta scale is not going to alter the position of minus infinity, and therefore will not effect the value of $\underline{c}i$. Next, if our scale transformation is given by [1.1] then our $\underline{\theta}a$ and

bi must follow this same transformation because they are both based on this same metric. Finally, we must remember that any transformation of our parameters must not change the probability of a correct response to an item given a particular level of ability. That is, a change of the numerical scale value attached to a particular level of ability does not alter the conditional probability of passing the item. In typical models of the type with which we are concerned here, this probability is a function of ai(Θa-bi). The item discrimination transformation is given by: $ai^* = ai / A$. The reason for this form is that for any admissible values of A, B, ai, bi, ci,

$$ai^*(\Theta a^*-bi^*) = ai(\Theta a-bi).$$

Because our probability values are functions of these quantities, they too remain unaltered.

In summary, the entire problem of converting item parameters to a common metric hinges on identifying the correct linear transformation of our comparison group scale. If the parameter estimates were error free, as in the example given in this chapter, the problem would have a simple solution: for any item find the linear transformation of the scale for one group that causes the ICC for that item to match the ICC for the same item in the other group. Once this transformation has been identified, all parameters could be transformed according to equations [1.2a-d].

Unfortunately the above procedure is not generally applicable. This is because our parameter estimates are not error free. Some of the difference between corresponding item parameters estimated from

independent samples may be explained in terms of differences in metric, while another portion of this difference may be due to error in parameter estimation. For any given problem, the exact contribution from each source is difficult to determine. Several approaches, however, have been suggested to identify the appropriate linear transformation when error of estimation is present. These techniques are described in the following Chapter.

CHAPTER 2

TECHNIQUES FOR TRANSFORMING PARAMETERS TO A COMMON METRIC IN ITEM RESPONSE THEORY

Having identified the basic problem in the previous chapter, this chapter will examine seven approaches that have been proposed to find the appropriate scale transformation that places two or more independently estimated sets of parameters on a common metric. A theoretical presentation, along with a disscussion of criticisms of each technique is given. These techniques can be roughly classified into three categories. The first category involves three approaches that rely on information supplied by the estimated b-parameters from each group. These approaches all find the transformation that equates the first two moments of the distribution of estimated $\underline{b}i$'s between groups. They differ in the way poorly estimated difficulty parameters are treated.

The second class of techniques incorporates test and item characteristic curves to estimate the equating constants necessary for transforming to a common metric. Stocking and Lord (1982) suggest a method that examines a weighted test characteristic curve, while the two methods suggested by Haebara (1980), and Segall & Levine (1983) examine weighted sums of squared differences between corresponding ICC's.

Finally, the last technique, suggested by Segall examines vectors of estimated item parameter differences for corresponding items from the

two groups. This technique finds the values of the equating constants that maximizes a criterion related to the likelihood of observing these vectors of item parameter differences.

Equating Metrics using the First Two Moments of the Distributions of Estimated Item Difficulties

If the difficulty parameters for our two groups of independent parameter estimates were measured without error, the task of finding a common metric would be greatly simplified. Remember that from eq. [1.2b] we have:

$$\underline{b}i^{*} = A \times \underline{b}i + B. \qquad [1.2b]$$

If we examined the $\underline{b}i$ values for any two items (say items 1 and 2), we would have a system of two linear equations with only two unknowns (A and B):

$$b1^{*} = A \times b1 + B$$

$$b2^{*} = A \times b2 + B$$

Solving these equations for our equating constants A and B would be a simple task. Unfortunately our $\underline{b}i$'s are not measured without error, so this approach would very likely produce poor estimates of our A and B.

Instead, let us examine an approach that incorporates information supplied by all the estimated $\underline{b}i$'s from both groups. The basic motivation for this approach stems from the premise that if both our comparison and base groups are on equivalent scales, then the mean and standard deviation (SD) of the estimated $\underline{b}i$'s should also be equivalent

across groups. Intuitively this approach has some appeal, for both, the mean and *SD aggregate over* individual $\underline{b}i$'s, allowing for errors of measurement contained in these estimates to cancel out (or so it is hoped). On the other hand, if our two sets of parameter estimates are not on equivalent metrics, we would not expect to observe equivalent means and SD's of the estimated $\underline{b}i$'s for the two groups. In this case, however, there exists a linear transformation of the theta scale for one group, that will equate the mean and SD of our estimated $\underline{b}i$'s for the two groups. Once we find this linear transformation, we can then apply the components (A and B) of this transformation to all the parameters of the comparison group (as indicated by eq. [1.2a-d]).

We may now formalize the problem under this approach as one of finding the linear transformation of the theta scale for the comparison group by equating the first two moments of the difficulty parameters across groups. We may further elaborate our goal as one of finding the values of A and B such that:

$$\bar{b}^{*}(\text{comp}) = \bar{b}(\text{base}) \qquad [2.1a]$$

and

$$SD^{*}(\text{comp}) = SD(\text{base}) \qquad [2.1b]$$

where

$$\bar{b}^{*}(\text{comp}) = \frac{\sum_{i}^{n}[A\underline{b}i(\text{comp}) + B]}{n} \qquad [2.1c]$$

and

$$SD^*(comp) = \sqrt{\frac{\sum_{i}^{n}[Abi(comp) + B - \bar{b}^*(comp)]^2}{n}} \quad [2.1d]$$

and where $\bar{b}(base)$ = mean of base group b-parameters, and

SD(base) = the standard deviation of the base group b-parameters.

We may simplify [2.1c] to obtain:

$$\bar{b}^*(comp) = A \times \bar{b}(comp) + B \quad [2.3]$$

and by substituting [2.3] into [2.1d], we can simplify and obtain:

$$SD^*(comp) = A \times SD(comp) \quad [2.4]$$

Now from eq [2.1b] and eq [2.4] we have:

$$A \times SD(comp) = SD(base) \quad [2.5]$$

Solving eq [2.5] for A, we obtain:

$$A = SD(base) / SD(comp) \quad [2.6]$$

And now, from eq [2.1a] and eq [2.3] we obtain:

$$\bar{b}(base) = A \times \bar{b}(comp) + B \quad [2.7]$$

Solving this equation for B:

$$B = \bar{b}(base) - A \times \bar{b}(comp) \quad [2.8]$$

and substituting eq [2.6] for A:

$$B = \bar{b}(base) - [SD(base)/SD(comp)] \times \bar{b}(comp) \quad [2.9]$$

Thus equations [2.6] and [2.9] specify the expressions for our equating constants. When these constants are applied to the scale of

our comparison group, the first two moments of the distributions of estimated item difficulties are equal to those of the base group. Once we obtain the values of our equating constants (A and B) from eq [2.6] and [2.9] respectively, we can transform all the parameters of the comparison group to the metric of the base group by use of equations [1.2a-d].

There is, however, a serious shortcoming with the procedure as outlined above. The shortcoming centers around our error of estimate for the b-parameters used to find our A and B. As pointed out previously, by basing our A and B on the mean and SD of our estimated item difficulties, we hope that errors of measurement contained in the difficulty parameter estimates cancel out. However, this may not happen. Poorly estimated difficulties may have a large influence on the sample moments, producing equating constants that are poor indicators of the transformation necessary to equate the two groups of parameters. Several "fix-ups" have been proposed to deal with the problem of the effect of poorly estimated difficulties on the sample moments. Two of these "fix-ups" are described in the following sections.

Difficulty Parameter Equating with Restricted Range of Discrimination and Difficulty Parameter Values

One way to reduce the effect of poorly estimated $\underline{b}i$'s on the computation of sample moments is to exclude items with extreme difficulty values (eg. $|\underline{b}i| > 3$). The error of estimate of the $\underline{b}i$'s

for these items is high relative to items with moderate difficulty values. Also, items with low discrimination values (a_i's) should be excluded from the computation of the sample moments (eg. $|a_i| < .15$). These items also have large sampling variances for the b_i's. The goal is to obtain a smaller set of better estimated b_i's for each group from which to compute our sample moments, as outlined in the previous section.

As might be anticipated, one of the major drawbacks of this approach is that it is hueristic in nature, offering no firm statement as to which items to exclude. Rather, we have only a few rules of thumb to follow. In an attempt to remedy this condition, and to control for the effects of the pooly estimated b_i's on the sample moments in a more systematic manner, Linn, Levine, Hastings and Wordrop (1980) suggest the procedure outlined in the following section.

Difficulty Parameter Equating using Weighted Moments

Linn, Levine, Hastings and Wardrop (1980) controlled the effects of poorly estimated b_i's by the use of weights that are inversely proportional to the estimated variance of the estimated item difficulties. (See equations 2.11a-b.) The weights are applied to the b-parameters for each group, producing weighted means and standard deviations from which our A and B are derived as outlined in the previous section. Thus, items with large standard errors of their b_i, would receive less weight in the computation of the means and SD's

relative to items with small standard errors of their $\underline{b}i$.

The weighted indices are computed as follows:

STEP 1: First item covariance matrices are computed by inverting the 3x3 information matrix for each item, for each group. Formulas for the elements of the information matrix are given in Lord (1980, p.191), for the three parameter logistic model. Thus, for each item, two covariance matrices are computed, one from parameter estimates of the base group, and the other from the comparison group parameter estimates.

STEP 2: Next, the diagonal element for the variance of the difficulty parameter is extracted from each pair of covariance matrices for corresponding items; one variance term coming from the base group covariance matrix, and the other term from the comparison group covariance matrix. The larger of the two variance estimates is used in computing the weight for that item. If we let Vi(base) and Vi(comp) be the estimated sampling variances of $\underline{b}i$(base) and $\underline{b}i$(comp) respectively, the weight for item i is:

$$Wi = \begin{cases} 1 \,/\, Vi(base) & \text{if } Vi(base) \geq Vi(comp) \\ 1 \,/\, Vi(comp) & \text{if } Vi(comp) \geq Vi(base) \end{cases} \qquad [2.10]$$

The effect of selecting the larger of the two variances was to give the greatest weight to those items that possessed relatively small estimated sampling variances in both groups. If the difficulty parameter was poorly estimated in either sample (base or comparison), then it would receive a small weight relative to a $\underline{b}i$ that was well estimated in both samples.

Notice, however that there may be problems with comparing the two estimates of the sampling variances Vi(base) and Vi(comp) at this point.

The value of the Vi's is dependent (in part) on the unit of the scale on which the $\underline{b}i$'s are measured. That is, we would expect that the choice of unit for either our base or comparison groups would effect the respective values of these estimated variances. (The exact nature of this relationship is given by eq [2.31b]) Since the two scales for our base and comparison groups are not on equivalent metrics at this point, comparisons of these sampling variances across groups may not be appropriate.

STEP 3: The next step involves using these weights to compute the weighted means and SD of the $\underline{b}i$'s for the base and comparison groups. The weighted means of the $\underline{b}i$'s are computed for each group as:

$$\bar{b}w(base) = \sum_{i}^{n}[Wi \times bi(base)] / k \qquad [2.11a]$$

$$\bar{b}w(comp) = \sum_{i}^{n}[Wi \times bi(comp)] / k \qquad [2.11b]$$

The weighted SD for each group is computed as:

$$SDw(base) = \sqrt{\frac{\sum_{i}^{n} Wi[\underline{b}i(base) - \bar{b}(base)]^2}{k}} \qquad [2.12a]$$

$$SDw(comp) = \sqrt{\frac{\sum_{i}^{n} Wi[\underline{b}i(comp) - \bar{b}(comp)]^2}{k}} \qquad [2.12b]$$

where

$$k = \sum_{i}^{n} Wi$$

STEP 4: Once the weighted means and SD's have been computed for the comparison and base groups using the above formulas, they can be

incorporated into formulas [2.6] and [2.9] just as their unweighted counterparts to find the values of our equating constants A and B.

The procedure described above attempts to control the influence of the sampling error of the bi's. It is interesting to note that although the error of estimate for the difficulty parameter for a particular item may be relatively high, we may still know a great deal about the shape of the item characteristic curve. Because the shape of the ICC is determined by the values of all three item parameters, a procedure that relies on the shape of the ICC's may be more informative then these procedures that examine only the bi's. In the next section, we turn to a class of techniques that use all three item parameters (ai, bi, and ci) in an attempt to find the linear transformation necessary to develop a common metric.

Sums of Squared Differences Between Estimated True Scores

Stocking and Lord (1982) suggest a technique that uses true scores to find the comparison group scale transformation. Each member of an arbitrarily selected group possesses an estimated true score. That is, an examinee, a, with ability Qa, has an estimated true score $\hat{\xi}_a$ defined by:

$$\hat{\xi}_a = \sum_i^n [Pi(Qa;ai\text{-}hat,bi\text{-}hat,ci\text{-}hat)] \qquad [2.13]$$

If two different calibrations of the same test resulted in parameter estimates that were based on comparable metrics, then we would expect the difference in true scores for examinee a, from these two calibrations to be small. If on the other hand, the parameter estimates for the two calibrations were not based on equivalent metrics, then we would expect to see a larger discrepancy in the two true score estimates, for examinee a, based on the two sets of item parameter estimates. These observations suggest the following approach. Utilizing our familiar terminology, we may represent the true score estimate for a member of our base group as:

$$\hat{\xi}_a(\text{base}) = \sum_i^n [Pi(Qa(\text{base});ai\text{-hat}(\text{base}),bi\text{-hat}(\text{base}), ci\text{-hat}(\text{base})] \qquad [2.14a]$$

Notice that this estimate incorporates parameter estimates (ai-hat, bi-hat, ci-hat) obtained from our base group calibration. We may specify an alternative true score estimate for members of our base group as:

$$\hat{\xi}_a^*(\text{base}) = \sum_i^n [Pi(Qa(\text{base});ai\text{-hat}^*(\text{comp}),bi\text{-hat}^*(\text{comp}), ci\text{-hat}(\text{comp})] \qquad [2.14b]$$

where this estimate is computed from comparison group parameter estimates, that are transformed to the base group scale. (These transformations are given in eq [1.2a-b].) Thus, we would like to find the values of A and B that would minimize the difference $[\hat{\xi}_a(\text{base}) - \hat{\xi}_a^*(\text{base})]$. Stocking and Lord propose the following function to be

minimized:

$$F = (1/N) \times \sum_{a}^{N} [\hat{\xi}_a(\text{base}) - \hat{\xi}_a^*(\text{base})]^2 \qquad [2.15]$$

We wish to find values of A and B such that the average squared difference between the two true score estimates for members of our base group is a minimum. Ideally, to find the value of A and B that minimize [2.15] we would take the partial derivatives with respect to A and B, set these expressions equal to zero, and then solve for A and B. This approach, however is not possible in this instance because there is no closed form solution for A and B once the partials are set equal to zero. Thus, to find the values of our equating constants, an iterative numerical procedure must be used.

There are several observations that may help clarify this procedure. First, we should note that we are only dealing with true score estimates from members of our base group. True score estimates from members of the comparison group do not enter into the computations. Of course, the decision of which group is base versus comparison is arbitrary. Accordingly, the decision as to which group of true scores will be examined by this procedure is also arbitrary. The point to be stressed, however, is that only the true score estimates from one of the two groups are used.

A second observation worth noting is that in eq [2.14a-b] we acted as if we were using the true theta values rather than their estimated values. In practice, the true thetas are never known and the estimated ability parameters are used in their place.

A final observation might serve to clarify the role of our A and B in the minimization of F (eq [2.15]). Our A and B equating constants influence the values of ai-hat*(comp) and bi-hat*(comp) in expression [2.14b]. Their exact influence is specified by eq [1.2a-b].

Another way to conceptualize the function to be minimized by the current procedure is by: the squared difference between the test characteristic function for the base group, and the "transformed" test characteristic function for the comparison group, where this squared difference is weighted by the number of base group examinees occurring at each value of Θ. This conceptual approach is easily reconciled with the concept of squared differences between true scores by remembering that both true score, and a point along the test characteristic curve (TCC), may be expressed as:

$$\hat{\xi}_a = \text{True-Score}(\Theta a) = n \times \text{TCC}(\Theta a) = \sum_i^n \text{Pi}(\Theta a) \qquad [2.16]$$

(Where n represents the number of items in the test.) We can think of each Θa along the test characteristic curve as being weighted by the number of base group examinees with ability "Θa".

This latter conceptualization of the current approach helps bridge the gap between this technique and the next two methods described in the following section. In an attempt to find the proper scale transformation for the comparison group, the two techniques described in the following section examine the squared differences between corresponding <u>item</u> characteristic curves, rather than the squared

differences of the test characteristic curve. An additional distinguishing feature of the current method and the two that follow, is in the manner by which estimated θa's are used to weight these "squared differences." Whereas Stocking and Lord use estimated θ's from only the base group, both of the following techniques use information supplied by both groups, base and comparison, in deriving weights for developing the linear transformation necessary to form a common metric.

Squared Difference Between Corresponding Item Characteristic Curves

Haebara's Method:

As in each of the previous techniques, our task is to find the linear transformation of the comparison group scale that places the estimated parameters of this group and those of our base group on a common metric. Haebara (1980) suggests a method that finds this transformation by examining the sum (across items) of the weighted sum of squared differences between corresponding ICC's.

If our item parameters were measured without error, then using any item i, we could find the values of A and B, such that for every value of θ:

$$Pi(\theta(comp);ai(comp),bi(comp),ci(comp)) = Pi(\theta(comp) \times A + B;ai(base);bi(base),ci(base)) \quad [2.17]$$

Notice that [2.17] implies perfect equating. Since our item parameters

are not measured without error, we would not expect this relationship to hold exactly. Instead, we would like to find the values of A and B such that the equality in [2.17] holds as closely as possible across items in our test. This objective motivates the following specification for the criterion function. Let:

$$ERia(comp) = Pi(\Theta a(comp);ai(comp),bi(comp),ci(comp)) - Pi(\Theta a(comp) \times A + B;ai(base),bi(base),ci(base)) \quad [2.18]$$

Then one candidate for our criterion function would be:

$$Q(comp) = \sum_{i}^{n} \sum_{a}^{Nc} [ERia(comp)]^2 \quad [2.19]$$

That is, for each examinee in our comparison group, on each item, we examine the squared difference between two probabilities. One probability is obtained from the estimated item parameters for our comparison group (by way of the logistic or normal ogive models). The other probability is obtained from transforming Θa to the base group metric and employing our estimated base group parameters for that item. Finally, we sum the squared differences of corresponding probabilites across people in our comparison group, and then across items.

For practical reasons (dealing with computation time and computer storage), Haebara uses an approximation to the quantity in [2.19] that incorporates a relative frequency distribution of $\Theta(comp)$ rather than using each individual value. This relative frequency distribution h(comp) of $\Theta(comp)$ is constructed by dividing the range of $\Theta(comp)$ into

k small intervals with the midpoints Θj(comp) (where j=1,2,...k). Then minimizing Q(comp) is approximately equal to minimizing Qh(comp), that is define as:

$$Qh(comp) = \sum_{i}^{n} \sum_{j}^{k} [ERij(comp)]^2 \times hj(comp) \qquad [2.20a]$$

Notice however that this quantity is based on equating errors from members of our comparison group only. Haebara defines an analogous quaitity that expresses the contribution of our base group examinees to the equating criterion. This quantity is defined as:

$$Qh(base) = \sum_{i}^{n} \sum_{j}^{k} [ERij(base)]^2 \times hj(base) \qquad [2.20b]$$

where h(base) is the relative frequency distribution of our base group examinees and:

$$ERij(base) = Pi(\Theta j(base);ai(base),bi(base),ci(base)) - Pi([\Theta j(base)-B]/A;ai(comp),bi(comp),ci(comp)) \qquad [2.21]$$

Notice that if we allow the quantity:

$$\Theta^*(base) = \Theta(comp) \times A + B \qquad [2.22a]$$

to represent the transformation of our comparison group ability parameter to the base group metric, then solving [2.22a] for Θ(comp), we obtain:

$$\Theta^*(comp) = [\Theta(base) - B] / A \qquad [2.22b]$$

which represents the transformation of our base group ability parameters to the comparison group metric, as implied by [2.21]. Then Haebara suggests:

$$Q^* = Qh(comp) + Qh(base) \qquad [2.23]$$

as the final form of the criterion function. Our task has now been reduced to one of finding the values of our equating constants A and B that minimize Q^*. As in the case of the technique suggested by Stocking and Lord (1982) there is no closed form solution for A and B. Thus, to find the value of A and B that minimize [2.23] we must employ an iterative numerical procedure.

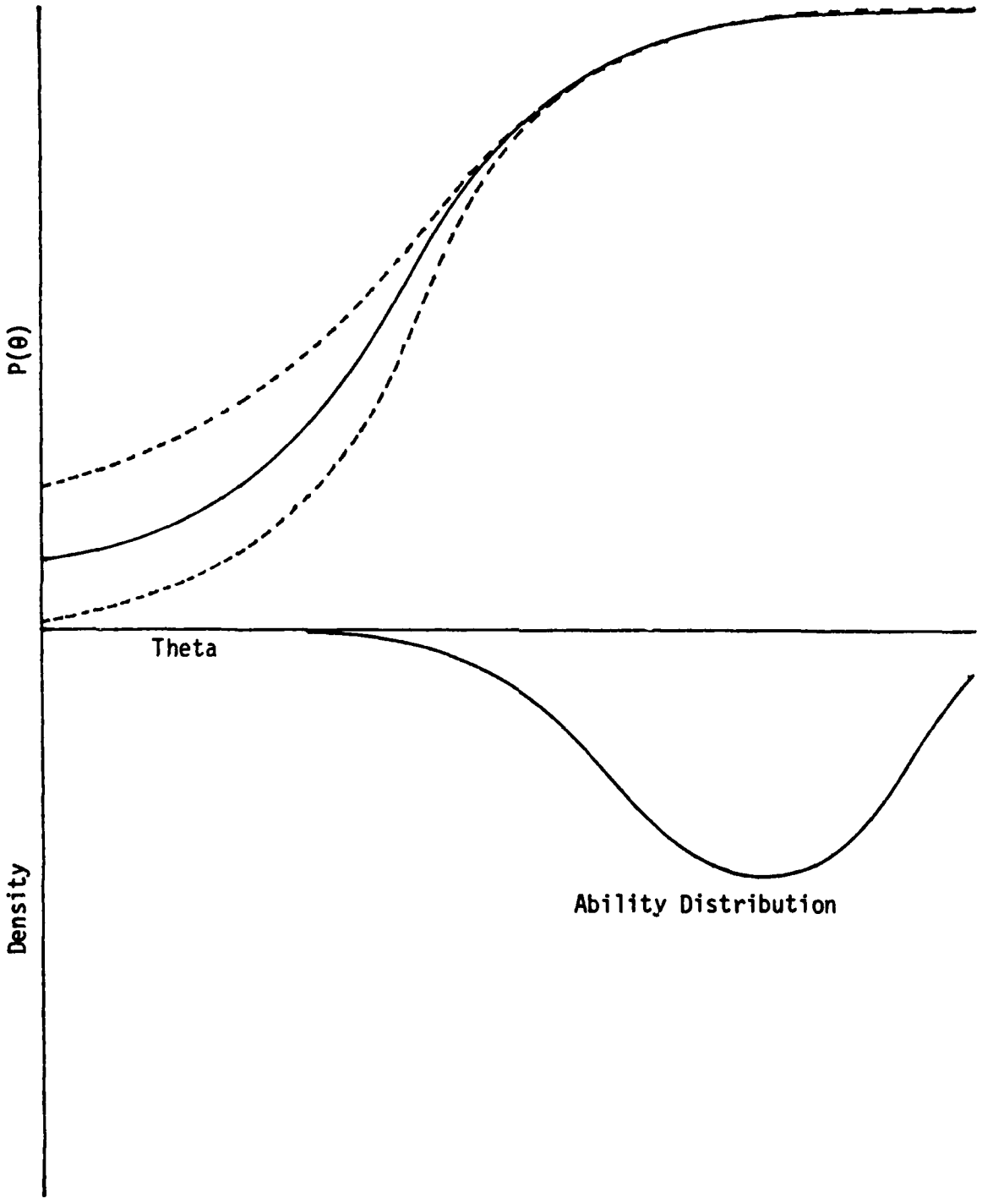
One criticism with Haebara's technique deals with how the squared differences between corresponding ICC's are weighted. To understand fully the rationale behind this criticism, a slight diversion might be helpful.

Figure 5 displays an arbitrary item characteristic curve (solid line) and a hypothetical distribution of examinees (represented by the u-shaped curve along the base line). If we used a sample from this distribution of examinees to estimate the shape of the ICC in Figure 5, we could very likely end up with an estimated curve represented by either of the dashed lines in Figure 5. Notice that where we have the greatest number of examinees, the agreement between the true ICC (solid line) and our estimates is very close. Where we have very few, or no examinees, the discrepancy between our estimated and true ICC may be very large. In general, we can place a great deal of confidence in the shape of a particular segment of an ICC, if in the region of that segment we have a substantial number of examinees. Conversely, if along a particular segment of an ICC we have very few examinees, we should place very little confidence in its estimated shape.

We may generalize this argument to our present situation, where we have two estimates for every ICC, each estimated from a potentially

Figure 5

Relation of True ICC and Hypothetical Estimates of the ICC with Distribution of Ability

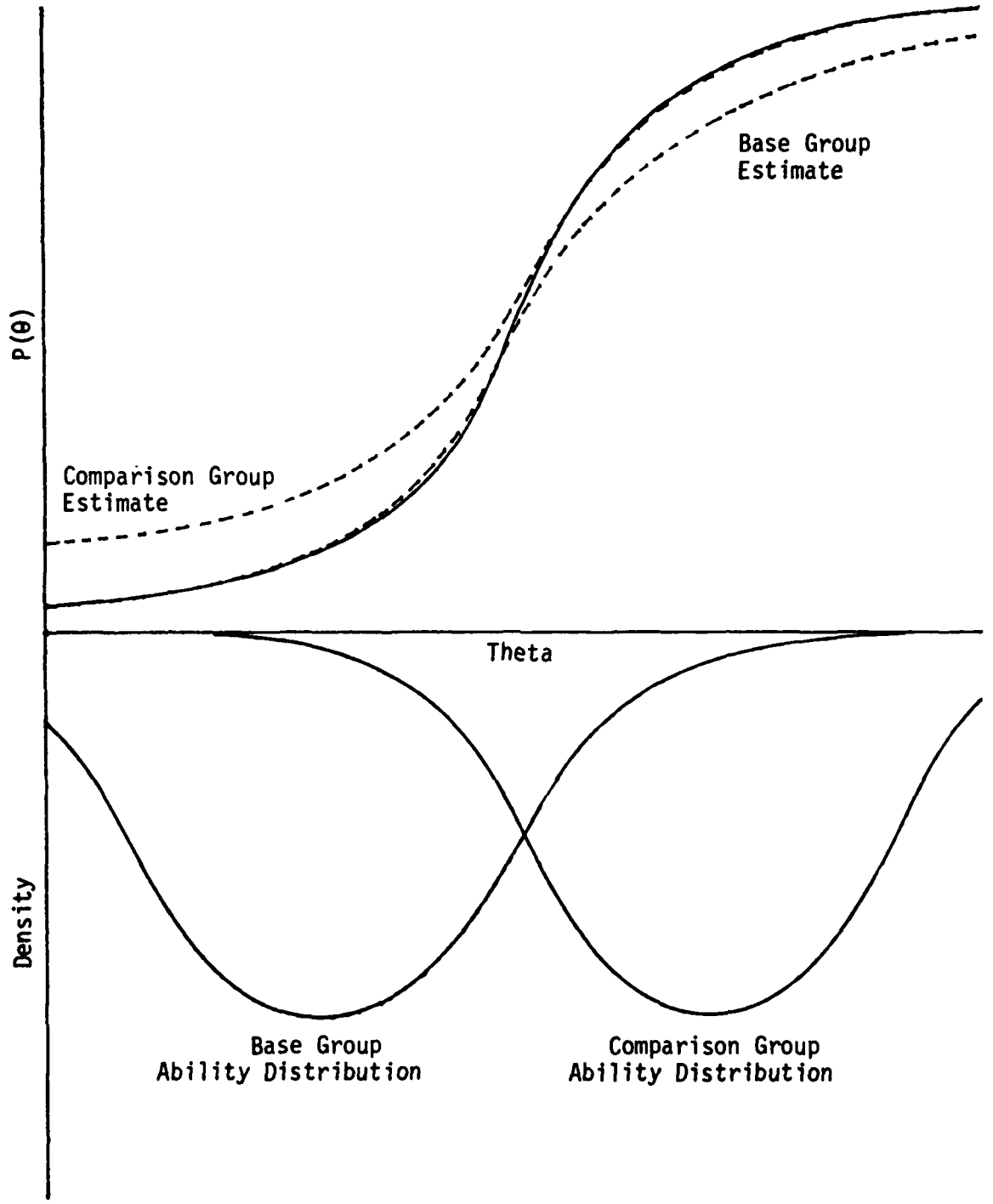


different distribution of ability. A hypothetical example is given in Figure 6. In this illustration we have two distributions of ability as indicated by the two u-shaped curves along the base line. From each sample obtained from these hypothetical distributions, we derive an estimate for our true ICC. (The true ICC is indicated by the solid line, and each of the estimates, from each sample, by the dashed lines.) Notice, as we might expect, that the base group estimate is very close to the true ICC along the range where there are a large number of base group examinees. Similarly, the estimated comparison group ICC is very close to the true ICC along the range where there are a large number of comparison group examinees. Also, as we might have anticipated, the discrepancy between the true and estimated ICC can be very large in areas where there are very few or no examinees, from the corresponding group, from which to derive the estimate.

The weighting scheme suggested by Haebara (1980) weights those segments of the estimated ICC differences in accordance with the relative frequency of examinees falling in the region. Returning to Figure 6, the weight function for the base group h(base) would weight the squared differences between one curve that was estimated very well along this range (the base group estimate) and one curve that was estimated very poorly along this range (the comparison group estimate). An analogous point can be made concerning the weight function for the comparison group, h(comp). The point to be stressed here is that a substantial portion of the weighted squared differences between corresponding ICC's may be due to error of estimation when a weighting scheme such as the one suggested by Haebara is employed.

Figure 6

Relation of True ICC and Two Independent Estimates of the ICC With Two Different Distributions of Ability



Segall and Levine (1983) suggest a weighting scheme that attempts to eliminate the above criticism. Their approach is outlined in the following section.

Segall and Levine Method:

The quantity to be minimized by the technique suggested by Segall and Levine (1983) is Qf, define as:

$$Qf = \sum_{i}^{n} \sum_{j}^{k} [ERij]^2 \times fj^* \qquad [2.24]$$

where:

$$ERij = Pi(\Theta j(comp);ai(comp),bi(comp),ci(comp)) - Pi(\Theta j(comp) \times A + B;ai(base),bi(base),ci(base)) \qquad [2.25]$$

fj^* represents a new weight function which is obtained in the following manner. First the relative frequency distribution of Θ(comp) is constructed by dividing the range of Θ(comp) into k small intervals with the midpoints Θj(comp). Next these relative frequencies are transformed to relative proportions to produce fj(comp). Then, fj(comp) is transformed to the scale of our base group (using our equating constants A and B) and fj(base) is computed using these transformed cut-points on our distribution of base group examinees. Our complete weight function is then computed as:

$$fj^* = \sqrt{[fj(comp) \times A + B] \times [fj(base)]} \qquad [2.26]$$

Returning to Figure 6, notice that our new weight function will be largest over the range where the overlap of the two estimated distributions of ability is the greatest. The weight function will be

smallest, or zero, where we have little or no overlap of our two estimated distributions of ability. This weighting scheme places the heaviest emphasis on that portion of the squared difference between corresponding ICC's that are relatively well estimated in both groups. That portion of the squared difference between corresponding ICC's that is computed from an ICC segment which is poorly estimated for either group, receives a small or zero weight.

As in the case of the two previous methods there is no closed form expression for the minimization of Qf in eq [2.24], so an iterative numerical procedure must be used.

The method described in the following section utilizes a somewhat different approach to control the influence of parameter sampling error on the estimation of our equating constants. This technique, suggested by Segall(1982), employes estimated covariance matrices for our item parameters in the framework of a maximum likelihood estimation procedure.

Estimation of Equating Constants Using Vectors of Item Parameter Differences

In this section a method is introduced by adopting maximum likelihood estimation concepts to the problem of estimating the equating constants. A heuristic discussion will be used to review the reasoning that led to the method. Of course, the heuristic argument is not

essential. First consider one item and the two vectors of item parameter estimates

$$\alpha_i(\text{base}) = \begin{bmatrix} a_i(\text{base}) \\ b_i(\text{base}) \\ c_i(\text{base}) \end{bmatrix}$$

$$\alpha_i(\text{comp}) = \begin{bmatrix} a_i(\text{comp}) \\ b_i(\text{comp}) \\ c_i(\text{comp}) \end{bmatrix}$$

Since these vectors are maximum likelihood estimates from large samples, they will be approximately multivariate normal. Since they are estimated from different samples they will be independent, and any linear combination of them will be multivariate normal.

Let Ao and Bo denote the "true" equating constants, i.e. the unique pair of constants that transforms the ability scale of the comparison group to the ability scale of the base group after the two scales have been specified. For each A and B the vector

$$v_i = v_i[A,B] = \begin{bmatrix} a_i(\text{base}) \\ b_i(\text{base}) \\ c_i(\text{base}) \end{bmatrix} - \begin{bmatrix} a_i(\text{comp}) \,/\, A \\ b_i(\text{comp}) \times A + B \\ c_i(\text{comp}) \end{bmatrix} \qquad [2.27]$$

must be multivariate normal because it is a linear combination of multivariate normal vectors.

The covariance matrix of v_i can be easily determined from the covariance matrices of $\alpha_i(\text{base})$ and $\alpha_i(\text{comp})$. Maximum likelihood

estimation theory can be used to estimate the covariance matrices of each of the component vectors. For each possible value of the equating constants vi will be multivariate normal with an approximated covariance matrix C(A,B). (In fact, C(A,B) is independent of B, but that fact is not needed here.)

If only the expectation of the random vector vi were known, its multivariate normal density could be specified. If A and B are equal to Ao and Bo respectively, and the maximum likelihood estimates are based on large enough samples to be considered unbiased, the expectation $E[\alpha_i(\text{base})]$ will equal the linearly transformed $E[\alpha_i(\text{comp})]$, and vi will have expectation equal to zero.

To summarize, for each A and B the hypothesis A=Ao and B=Bo implies that vi is multivariate normal with zero expectation and specified covariance matrix. The hypothesis implies a specific formula for the multivariate density of vi. If the estimates for different items were independent, then the joint distribution of all the vi would also be multivariate normal with density

$$L[v1(A,B),v2(A,B),\ldots,vn(A,B)\,|\,A=Ao \text{ and } B=Bo] = \prod_i L[vi(A,B)\,|\,A=Ao \text{ and } B=Bo] \qquad [2.28]$$

Unfortunately the estimates for different items are not completely independent when item and person parameters are estimated simultaneously. Minor dependencies are expected and observed when the same group of subjects are used in the estimation of each item. The method introduced in this section ignores these interdependencies and

treats formula [2.28] as the joint conditional density of the vi. An estimate of Ao and Bo is obtained by maximizing this joint density.

This admittedly heuristic argument led to the formulation of a new technique which in fact performs quite will. The vectors of item parameters for corresponding items are treated as if they were independent multivariate normal vectors with covariance matrices specified by inverting estimated information matrices. A and B are estimated by maximizing the joint density under the hypothesis that A=Ao and B=Bo. It will be seen that inspite of the correlations between the parameters of different items, that the method performs quite well. Some details on the implementation of the method follow.

For the present let us develop our criterion on a basis of one item only. The generalization to an n-item test is straight forward and will be discussed later in this section. For the moment, our task is to make explicit a probability function for the vi where the estimated vectors

$$\begin{bmatrix} ai(base) \\ bi(base) \\ ci(base) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} ai(comp) \;/\; A \\ bi(comp) \times A + B \\ ci(comp) \end{bmatrix}$$

are independently sampled from a multivariate normal population with known covariance matrix. That is, we seek a formula for

$$\text{Prob}\{vi(A,B) \mid A=Ao, B=Bo\} \qquad [2.29a]$$

where Ao and Bo are the true equating constants. The more compact and suggestive notation

$$L(vi \mid A=Ao, B=Bo) \qquad [2.29b]$$

will be used to denote this multivariate probability density. Because

the item parameters are maximum likelihood estimates of essentially the same parameters in each group, we can regard the vi as normally distributed with expectation equal to zero. That is, because the expectation of each vector of item parameters is their true value, and because their true values are the same for corresponding items i, then the expectation of their difference is a zero vector.

Next, we would like to examine the sampling variance of our estimated item parameters. For the moment, let us concern ourselves with a single vector of parameter estimates. Maximum likelihood theory specifies that the variance - covariance for these estimates may be obtained from the inverse of the information matrix (I). When the ability parameters are known, formulas for the elements of the information matrix are given in Lord (1980, p.191). Thus the sampling variance-covariance (Ci), for item i, of our item parameter estimates may be obtained from:

$$C = I^{-1} = \begin{bmatrix} Iaa & Iab & Iac \\ Iba & Ibb & Ibc \\ Ica & Icb & Icc \end{bmatrix}^{-1} \qquad [2.30]$$

We will have two covariance matrices for each item. One covariance matrix from our base group parameter estimates and the other from our comparison group parameter estimates. We may represent these as:

$$C(base) = I^{-1}(base) = \begin{bmatrix} Caa(base) & Cab(base) & Cac(base) \\ Cba(base) & Cbb(base) & Cbc(base) \\ Cca(base) & Ccb(base) & Ccc(base) \end{bmatrix} \qquad [2.31a]$$

$$C(comp)^{*} = I^{-1}(comp) = \begin{bmatrix} Caa(comp)/A^{2} & Cab(comp) & Cac(comp)/A \\ Cba(comp) & Cbb(comp)xA^{2} & Cbc(comp)xA \\ Cca(comp)/A & Ccb(comp)xA & Ccc(comp) \end{bmatrix} \quad [2.31b]$$

Notice that certain values of the elements of the covariance matrix for the comparison group, in eq [2.31b], depend on the value of our equating constant A. This is because we are transforming the metric of the comparison group to that of the base group, and we would expect this transformation to have some effect on our variance - covariance elements which were computed in our original comparison group metric.

Next, because our two estimates of parameters for an item are independent (each coming from a different group), the sampling variance of our vectors of parameter differences (vi) is equal to:

$$Cdi = Ci(base) + Ci^{*}(comp) \quad [2.32]$$

Finally to specify the criterion based on eq [2.29] we can examine the surface of a tri-variate normal density function, with N[0, Cdi]:

$$fi(vi|A,B) = (2xpie)^{**}(-3/2)\ |Cdi|^{**}(-1/2)\ exp(-cs/2) \quad [2.33]$$

where:

$$cs = xi'\ Cdi^{-1}\ xi \quad [2.34]$$

and:

$$xi = vi - E(vi) = vi - 0 = vi \quad [2.35]$$

Thus [2.33] gives us a criterion related to the likelihood of obtaining a single vector of parameter differences for given values of A and B.

To obtain the analogous quantity that incorporates information supplied by all the items in our test, we formulate the following objective function,

$$L(v1,v2,v3,.....vn|A=Ao \ \& \ B=Bo) = \prod_{i}^{n} f(vi|A,B) \qquad [2.36]$$

To find the values of A and B that maximize [2.36] an iterative numerical procedure is employed.

The next two chapters present a comprehensive comparison of the techniques outlined in this chapter. The relative abilities of these techniques to estimate accurately the linear transformation necessary to develop a common metric are assessed. This assessment involves, both simulated and real data, covering a variety of conditions.

CHAPTER 3

ASSESSMENT OF EQUATING TECHNIQUES USING SIMULATED DATA

This chapter presents a study whose objective is to assess the ability of each technique described in the previous chapter to accomplish its intended goal: to transform two sets of parameters to a common metric. To accomplish this assessment, two different approaches were taken.

The approach described in this chapter involves the use of simulated responses based on a simulated test and several different distributions of examinees. The use of simulated data provides greater control and knowledge concerning the true relation between the two sets of independently estimated parameters. Because the true relation between the two sets of parameters is known with simulated data, firm statements can be made concerning the ability of each technique to recover this relation. The main problem, however, with simulated data, is that it is based on a model whose assumptions are almost certainly violated to varying degrees by real people answering real items.

As an answer to this criticism, Chapter 4 presents an approach that examines the relative ability of the equating techniques to recover the proper transformation using "real" data. "Real" is used here in the sense that the data are actual responses to items on an actual test. The obvious drawback of this approach is that the true relation between the two sets of estimated parameters is not known. This presents a special problem for evaluating the accuracy of the estimated equating

constants. This problem, along with one solution, is presented in Chapter 4.

Experimental Design using Simulated Data

Unfortunately, there are an infinite number of simulated tests and simulated distributions of examinees that could be selected for inclusion in this study. Because it is feasible to examine only a few different simulated tests and distributions of examinees, we should select these carefully. First, it would be desirable to structure the test as closely as possible to the type of test that we would find in practice. Similarly, the distributions of ability should also be modeled after the types of distributions commonly observed. Modeling our simulated test parameters and item responses as closely as possible to actual tests makes generalization to "real" data easier. The design presented below attempts to specify values of these parameters that are similar to those found in many applied testing situations.

The Test

Table 3 lists the item parameters for a 60 item test used to generate the dichotomous item responses. The a-parameters of this test were specified by sampling numbers from a uniform distribution in the interval [.3, 1.4]. The b values were sampled from a uniform distribution in the interval [-3., +3.], and the c-parameters were drawn from a uniform distribution in the interval [.11, .33]. These parameters identify the items included in tests of three different

Table 3

Item Parameters for Simulated Test
Used to Create Simulated Response Vectors

ITEM	a	b	c
1	.7948	-2.8090	.2373
2	.3031	2.6931	.2594
3	1.0263	-.8141	.1978
4	.8035	-2.4044	.1300
5	.5906	-1.3875	.2613
6	.8063	1.8960	.1477
7	1.2265	-.6636	.2661
8	.3778	1.6944	.2494
9	1.2698	1.8150	.2885
10	.8935	.9261	.1688
11	.7382	.8374	.3083
12	1.3302	-1.8851	.2026
13	1.2925	2.9963	.1886
14	1.3972	1.4620	.2623
15	1.0194	1.2102	.1791
16	1.3878	.3561	.1377
17	1.0037	2.2912	.1355
18	.7901	.1613	.1786
19	.7742	-1.3758	.3221
20	1.0598	1.4167	.2454
21	.9478	2.8291	.3085
22	.9324	-1.5024	.1467
23	1.2279	2.8816	.3175
24	1.2978	-1.3201	.2998
25	.9555	.8139	.2150
26	1.0750	-.7296	.2105
27	.9491	1.6064	.3189
28	.6723	-1.0248	.3271
29	.4617	2.0528	.2669
30	.5510	1.7964	.3047
31	1.1574	.8547	.2619
32	.9271	.2455	.3261
33	.5457	-1.4978	.2188
34	1.3730	2.3566	.2960
35	.9273	.9723	.2323

(Table 3 continued)

ITEM	a	b	c
36	1.3273	2.6651	.2454
37	.5960	2.5812	.2540
38	1.3299	2.8950	.1447
39	1.2710	1.4454	.2847
40	.8222	-.6592	.2898
41	.6442	-2.9478	.2106
42	.4047	-2.5030	.2071
43	.5023	-1.5370	.2074
44	.7106	-1.7613	.2441
45	.7164	2.4118	.2937
46	.9033	-1.2684	.2822
47	.9312	-.1995	.2227
48	1.0233	1.3889	.2237
49	1.1059	2.8210	.2981
50	.6400	.8782	.1829
51	.8730	-.5741	.3145
52	.4710	-.7691	.2760
53	1.2326	-2.0874	.2392
54	1.3953	-1.5240	.1539
55	.7350	.5290	.1776
56	1.2183	-.6919	.2190
57	.4101	-.1285	.1582
58	1.2872	1.0834	.1707
59	1.2091	-1.5559	.3179
60	1.2616	-1.9413	.1828

lengths. In one set of conditions, Conditions 1 through 5 (see Table 4), parameters for all 60 items were used to generate dichotomous responses for a test of length 60. For Condition 6, the first 30 items parameters listed in Table 3 were used to generate responses for a test of length 30. And finally, in Condition 7, the first 15 items listed in Table 3 were used to generate responses for a test of length 15.

Ability Distributions

Several pairs of base group - comparison group ability distributions were examined. These pairs constitute the basic conditions of the simulation portion of this study.

In each condition, the mean and SD of the distribution of ability for the base group remained unchanged. These distributions were normal, with mean equal to zero, and standard deviation equal to one. In each condition, base group ability parameters were sampled from a normal [0,1] distribution. Because these sampled values were to be treated as true parameters, they were transformed to have zero mean and unit variance. (This transformation is similar to a z-score transformation.)

Several different comparison group ability distributions were specified, one for each condition listed in Table 4. Each of these distributions were generated by sampling values from a normal distribution. For the Conditions 1 through 7, the comparison and base group distributions differed by varying amounts. Table 4 lists the means and standard deviations for the ability distributions used in each condition. Notice that the mean and standard deviation are constant across conditions for the base group (with mean equal to zero and SD

Table 4

Summary of Simulated Conditions

Condition	Number of Subjects	Number of Items	Base Group		Comp Group		Scale Transformation	
			Mean	SD	Mean	SD	A	B
1	1000	60	0.0	1.0	-0.5	.80	.80	-0.5
2	500	60	0.0	1.0	-0.5	.80	.80	-0.5
3	250	60	0.0	1.0	-0.5	.80	.80	-0.5
4	500	60	0.0	1.0	0.0	.80	.80	0.0
5	500	60	0.0	1.0	-1.0	.80	.80	-1.0
6	1000	30	0.0	1.0	-0.5	.80	.80	-0.5
7	1000	15	0.0	1.0	-0.5	.80	.80	-0.5

equal to one). On the other hand, the mean and standard deviation for the comparison group did not necessarily remain constant from one condition to the next.

Samples of several different sizes were generated: 250, 500, and 1000 simulated examinees. In each condition the number of comparison group examinees equaled the number of base group examinees.

Generation of Item Responses

Data, or item responses for these simulated examinees were generated in accordance with Birnbaum's (1968) three parameter logistic model (see Table 1). Note that once the item parameters are specified (Table 3) the probability of a correct response to an item is solely a function of examinee ability. As examinee ability increases, so does the probability of a correct response. The probability of a correct response can be used to generate an observable dichotomous right-wrong response by comparing it to a uniformly distributed ramdom number between 0 and 1. A response was coded as correct when its associated probability was greater than the random number, and incorrect when it was less.

Using this procedure 14 data sets of dichotomous responses were generated, two data sets per condition (Table 4). Each data set contained the equivalent of N examinees answering an n item test. The item parameters from Table 3 along with person parameters sampled from a normal distribution were used in the above procedure to generate the data sets.

Estimation of Item and Person Parameters

The LOGIST computer program (Wood, Wingersky & Lord, 1976) was used to estimate all person and item parameters for each data set described above. Briefly:

> Given the responses of a group of examinees to a set of items, the computer program, LOGIST, has been developed to estimate the item characteristic curve parameters for each item and the ability of each examinee in terms of Birnbaum's three parameter logistic model. The parameters are estimated by a method analogous to the maximum likelihood method described by Lord (1968) with the likelihood function modified to handle omits. (Wood, & Lord, 1976, p.1)

Of special interest, for our purpose is the method LOGIST uses to specify the origin and unit of the measurement scale. Remember that these are arbitrary, and both person and item parameters are unidentifiable until these have been specified. LOGIST selects the unit and origin in such a manner that the final theta estimates ($\hat{\theta}$) have a mean of zero and a standard deviation of one, for all estimated thetas inside the range -THLIM to +THLIM. THLIM is either specified by the user or the default of 3.0 is used.

The method LOGIST uses to select the unit and origin was an important consideration in the selection of criteria used to judge the relative ability of the equating techniques to transform all parameters to a common metric. In a following section, several such criteria are discussed.

Estimation of Equating Constants

Each of the seven approaches described in Chapter 2 was used to

estimate the A and B equating constants for each condition listed in Table 4. For each of the seven conditions in Table 4 there was an associated pair of base group - comparison group estimated parameter sets from LOGIST. These base and comparison group parameters were input into each of the seven equating techniques. Notice that within each condition, each equating technique used as data the same estimated base and comparison group parameters. To obtain the estimated equating constants, computer programs were written for each of the seven techniques. These programs are listed in the APPENDIX.

Criteria for Recovery of the Equating Constants

The least complicated criterion for determining how closely each technique recovered the true transformation is to compare the estimated equating constants with the true values of these constants. The true value of these constants, for each condition, are listed in the last two columns of Table 4. Notice that the true values of the equating constants listed in the last two columns of table 4 are identical to the mean and SD of the comparison group distributions, listed in the previous two columns of Table 4. This relation was anticipated on the basis of the method used by LOGIST to specify the unit and origin of the theta metric. Remember that to generate a normal distribution with mean equal to B and a SD equal to A, one can take the values from a normal [0,1] distribution and apply the following linear transformation:

$$\theta j^{*} = \theta j \times A + B \qquad [3.1]$$

We have the identical situation with our LOGIST estimated person parameters. Remember that LOGIST fixes the mean of the estimated thetas to zero and their standard deviation to one. Thus the parameters for the base group will automatically be set to their original metric. The estimated thetas for the comparison group, however, will also have a mean of zero and a standard deviation of one. Then, it follows that the A and B components of the linear transformation used to set the estimated comparison group parameters to their original metric will correspond exactly to the true SD and mean, respectively, of the comparison group distribution for that condition. These values are listed in Table 4.

To judge the accuracy of each of the seven approaches, a comparison of the estimated equating constants with the true values can be made. The technique whose estimates come the closest to the true A and B can be ranked highest; the technique whose estimates come next closest, can be ranked second, etc.

One drawback with this approach is exemplified by the situation in which the estimated A constant of one equating technique is closest to the true A, while the estimated B constant of another equating technique is closest to the true B. Thus, in addition to examining the size of the difference between our true and estimated constants for each technique, it would be desirable to have an additional criterion that incorporated the effect of errors in both, the A and the B constants simultaneously.

One such criterion might involve the root mean squared error (RMSE) difference between the true thetas and the estimated comparison group

thetas, after equating. That is, from each of the seven equating techniques, under each condition, we obtain values of our A and B equating constants. From eq [1.2d], we can use these values to transform the estimated thetas of our comparison group. Accordingly one criterion would be:

$$RMSE_{tc} = \sqrt{\frac{1}{N}\sum_{a}^{N}\left[\theta_{ac(comp)} - (\hat{A}_{tc} \times \hat{\theta}_{ac(comp)} + \hat{B}_{tc})\right]^2} \qquad [3.2]$$

where the subscript t represents equating technique; c represents Condition; and N equals the number of subjects in the comparison group. The transformed thetas in the above RMSE belong to the comparison group only. If the estimated equating constants are close to their true values, we would expect the RMSE to be relatively small. On the other hand, we would expect a relatively large value of the RMSE for poorly estimated equating constants.

Notice that the size of the RMSE described by eq [3.2] is influenced by two factors. First, the size of each RMSE is influenced by error in the estimation of the comparison group thetas. For poorly estimated thetas we would expect a relatively large RMSE. For good estimates of these thetas we would expect to see a small RMSE. Second, the RMSE in eq [3.2] is also influenced by the error in the estimated equating constants. This is the portion of the RMSE that is of primary interest for our purposes. One possible improvement over the criterion given by [3.2] would be an index that was influenced only by the errors in the estimated equating constants. That is, it would be desirable to have an

index that was influenced only by errors due to equating and not by errors of estimate in the comparison group thetas. This desire motivates the following specification for a criterion that can be used to judge the accuracy of the equating constant estimates in each condition:

$$RMSE_{tc} = \sqrt{\frac{1}{N}\sum_{a}^{N}\left(\{\hat{A}_{tc}[(\theta_{ac}-B_c)/A_c] + \hat{B}_{tc}\} - \theta_{ac}\right)^2} \qquad [3.3]$$

Notice that the transformation in the inner brackets represents one that will give the true comparison group thetas a mean of zero and SD of one (just as they would have if LOGIST had estimated these parameters without error). This quantity is then transformed to the base group metric using the estimated constants. Finally the squared difference between the transformed theta and its true value is obtained and summed across examinees, to obtain the final RMSE. Notice that this index uses only true theta values, and thus avoids the problem associated with sensitivity to errors in person parameter estimation. The index given in eq [3.3] was computed for each technique in each of the seven conditons. Values of this index are listed in Tables 6 through 12.

Results

Tables 6 through 12 summarize the results of the simulated portion of this study. Each table lists the results of one condition given in Table 4. For each technique the estimated equating constants are given, along with the RMSE given by eq [3.3].

Table 13 list the average of the RMSE values across all seven

conditions for each technique. These averages were computed by taking the RMSE values listed in Tables 6 through 12, and averaging the RMSE values across the seven conditions, for each technique.

Table 5

Key for Tables 6 through 13

TECHNIQUE	Description
ALL b's	b-parameter equating (using all bi's)
SELECTED b's	b-parameter equating (using well estimated bi's)
WEIGHTED b's	b-parameter equating (using weighted bi's)
TRUE SCORE	True score equating (Stocking & Lord)
ICC (H)	ICC equating (Haebara)
ICC (S/L)	ICC equating (Segall & Levine)
MLE	MLE equating based on vectors of item parameter differences

Table 6

(Condition 1)

A = .80 B = -.50

NUMBER OF SUBJECTS

BASE GROUP: 1000 COMPARISON GROUP: 1000

NUMBER OF ITEMS: 60

TECHNIQUE	A	B	RMSE-COMP
ALL b's	.3679	-.5940	.4419832
SELECTED b's	.7456	-.6102	.1228830
WEIGHTED b's	.7797	-.5707	.0735640
TRUE SCORE	.7491	-.5191	.0543682
ICC (H)	.7695	-.5238	.0387308
ICC (S/L)	.7557	-.5257	.0512494
MLE	.7758	-.5025	.0242760

Table 7

(Condition 2)

A = .80 B = -.50

NUMBER OF SUBJECTS

BASE GROUP: 500 COMPARISON GROUP: 500

NUMBER OF ITEMS: 60

TECHNIQUE	A	B	RMSE-COMP
ALL b's	.1984	-.1578	.6916400
SELECTED b's	.8415	-.4719	.0500991
WEIGHTED b's	.7937	-.4246	.0756928
TRUE SCORE	.7249	-.4560	.0869498
ICC (H)	.7700	-.4637	.0470628
ICC (S/L)	.7236	-.4733	.0808992
MLE	.7708	-.4533	.0550536

Table 8

(Condition 3)

A = .80 B = -.50

NUMBER OF SUBJECTS

BASE GROUP: 250 COMPARISON GROUP: 250

NUMBER OF ITEMS: 60

TECHNIQUE	A	B	RMSE-COMP
ALL b's	.9640	-1.5273	1.0402315
SELECTED b's	.8006	-.4035	.0964754
WEIGHTED b's	.9011	-.4419	.1164287
TRUE SCORE	.7738	-.4799	.0329480
ICC (H)	.8038	-.4359	.0642359
ICC (S/L)	.7254	-.4655	.0820452
MLE	.7816	-.4146	.0873095

Table 9

(Condition 4)

A = .80 B = 0.00

NUMBER OF SUBJECTS

BASE GROUP: 500 COMPARISON GROUP: 500

NUMBER OF ITEMS: 60

TECHNIQUE	A	B	RMSE-COMP
ALL b's	2.2061	-.5012	1.4914301
SELECTED b's	.8067	-.0186	.0197350
WEIGHTED b's	.8509	-.0159	.0532301
TRUE SCORE	.7797	.0095	.0224022
ICC (H)	.8101	.0234	.0254557
ICC (S/L)	.7777	.0092	.0241489
MLE	.8196	.0526	.0561538

Table 10

(Condition 5)

A = .80 B = -1.00

NUMBER OF SUBJECTS

BASE GROUP: 500 COMPARISON GROUP: 500

NUMBER OF ITEMS: 60

TECHNIQUE	A	B	RMSE-COMP
ALL b's	.3859	-2.5905	1.6433882
SELECTED b's	.8670	-.8234	.1888696
WEIGHTED b's	.8725	-.8123	.2012114
TRUE SCORE	.7645	-.8925	.1132213
ICC (H)	.8104	-.8636	.1367862
ICC (S/L)	.7612	-.8802	.1259345
MLE	.8050	-.8475	.1525574

Table 11

(Condition 6)

A = .80 B = -.50

NUMBER OF SUBJECTS

BASE GROUP: 1000 COMPARISON GROUP: 1000

NUMBER OF ITEMS: 30

TECHNIQUE	A	B	RMSE-COMP
ALL b's	.2734	-.0522	.6910274
SELECTED b's	.7294	-.4557	.0832805
WEIGHTED b's	.8112	-.4340	.0669083
TRUE SCORE	.6850	-.4549	.1234248
ICC (H)	.7520	-.4466	.0718140
ICC (S/L)	.7390	-.4478	.0802616
MLE	.7470	-.4416	.0787914

Table 12

(Condition 7)

A = .80 B = -.50

NUMBER OF SUBJECTS

BASE GROUP: 997 COMPARISON GROUP: 1000

NUMBER OF ITEMS: 15

TECHNIQUE	A	B	RMSE-COMP
ALL b's	.5796	-.6473	.2650133
SELECTED b's	1.0096	-.4619	.2129480
WEIGHTED b's	.8863	-.2355	.2781787
TRUE SCORE	.8776	-.5151	.0790203
ICC (H)	.9005	-.3897	.1491785
ICC (S/L)	.8600	-.3953	.1206400
MLE	.9383	-.3644	.1936142

Table 13

Root Mean Squared Errors
Averaged Over all Seven Conditions
For Each Equating Technique

TECHNIQUE	MEAN-RMSE
ALL b's	.8950
SELECTED b's	.1106
WEIGHTED b's	.1236
TRUE SCORE	.0732
ICC (H)	.0762
ICC (S/L)	.0807
MLE	.0925

CHAPTER 4

ASSESSMENT OF EQUATING TECHNIQUES USING REAL DATA

This section examines the relative ability of the equating techniques to estimate the transformation using "real" data. Real in this context means the data are responses to items made by real people on an actual test, as opposed to simulated data described in Chapter 3.

The major emphasis of this portion of the study is to examine the effect that naturally occurring violations to the assumptions of IRT have on the equating procedures. We suspect that such assumptions as local indepencence, fit of the three parameter logistic model, and unidimensionality are violated to some degree by real people answering real items. By examining the performance of the equating techniques using real data, we may gain some insight into the effect these naturally occurring violations have on the performance of the equating techniques.

Study I

Data

Data for these analyses were obtained from the Anchor Test Study (Bianchini & Loret, 1974) equating study files. Item response data from the Word Knowledge (50 items) and the Reading Comprehension sections (45

items) of form F of the Metropolitan Achievement Tests (Durost, Bixler, Wrightstone, Prescott, & Balow, 1970) were obtained. The Reading Comprehension and Word Knowledge sections were combined and treated as a single 95 item test in the analyses described below. The subjects consisted of 2000 5th and 6th grade, white and black examinees.

Assignment of Subjects into Base and Comparison Groups

The 2000 subjects were randomly assigned to the base and comparison groups. This assignment resulted in 1000 examinees in each group.

Note that this random assignment of subjects results in expected values of the equating constants of A=1 and B=0. This is because random assignment of subjects results in expected distributions of ability that are equivalent across groups. If the distributions are equivalent, then their first two moments are also identical. Remember that LOGIST sets the unit of the scale equal to the SD of the estimated thetas and the origin of the scale equal to the mean of the estimated thetas. If the expected values of the mean and SD are equivalent for the base and comparison groups then the linear transformation that places the comparison group scale on the metric of the base group is simply:

$$\theta^* = A \times \theta + B$$

where A=1 and B=0.

Estimation of Item and Person Parameters

The LOGIST computer program (Wood, Wingersky & Lord, 1976) was used to estimate all person and item parameters for each group independently.

Results

Each of the seven approaches described in Chapter 2 were used to estimate the A and B equating constants. The base and comparison group parameter estimates from LOGIST were used as input for each of the seven techniques. Table 14 lists the estimated equating constants from each of the seven techniques.

Notice that all the techniques did extremely well at estimating the scale transformation. The first observation to be made is one concerning the effect of naturally occuring violations of the IRT model. With these data, none of the techniques appeared to be adversly effected by violations to the model. This is evidenced by the close agreement of all the estimated equating constants to their expected values.

The observation that all the techniques did well is not surprising. In this condition, there were a relatively large number of subjects in each group, each subject answering a relatively long (95 item) test. Perhaps most importantly for the three b-parameter equating techniques was the heterogeneity of the samples used to estimate these item parameters. As a result of selecting two samples with wide range of abilities (and in part due to the test's suitability to these samples), all the b-parameters had small sampling errors. Thus the transformation based on these 95 well estimated b-values was very close to the expected transformation.

Because all the techniques did so well in estimating the transformation, it may be interesting to assess the performance of these techniques under less ideal condtions than the one discussed above. The study outlined in the following section examines the performance of the

Table 14

MAT

A = 1.00 B = 0.00

NUMBER OF SUBJECTS

BASE GROUP: 1000 COMPARISON GROUP: 1000

NUMBER OF ITEMS: 95

TECHNIQUE	A	B
ALL b's	1.0042	.0011
SELECTED b's	1.0042	.0011
WEIGHTED b's	1.0079	-.0160
TRUE SCORE	1.0154	-.0154
ICC (H)	1.0151	-.0136
ICC (S/L)	1.0148	-.0131
MLE	1.0157	-.0144

See key on p.62 for identification of techniques.

equating techniques when smaller samples and fewer items are used.

Study II

Data

Data for these analyses were again obtained from the Anchor Test Study (Bianchini, et al.) equating study files. In this study, only item response data from the Reading Comprehension section (45 items) of form F of the Metropoliton Achievement Tests (Durost, et al.) were used. A new sample of 1000 5th and 6th grade, white and black examinees was used. Note that the examinees used in these analyses were exclusive of those used in Study I.

Assignment of Subjects into Base and Comparison Groups

The 1000 subjects were randomly assigned to the base and comparison groups. This assignment resulted in 500 examinees in each group. Notice that this random assignment of subjects again resulted in expected values of the equating constants of A=1 and B=0.

Estimation of Item and Person Parameter

The LOGIST computer program (Wood, et al.) was used to estimate all item and person parameters for each group independently.

Results

Each of the seven approaches described in Chapter 2 were used to

estimate the A and B equating constants. The base and comparison group parameter estimates from LOGIST were used as input for each of the seven equating techniques. Table 15 lists the estimated equating constants from each of the seven techniques.

Notice, from Table 15, that although the estimated transformations are not as close to the expected values as were those from Study I, the estimates from the present study all appear in fairly close agreement. Again, the relatively good performance of the three b-parameter techniques is most likely a result of the heterogeneous samples used to estimate these values.

The implications of the finding of these results along with those of the simulation portion of this study are discussed in detail in the following Chapter.

Table 15

MAT

A = 1.00 B = 0.00

NUMBER OF SUBJECTS

BASE GROUP: 500 COMPARISON GROUP: 500

NUMBER OF ITEMS: 45

TECHNIQUE	A	B
ALL b's	.9227	-.0058
SELECTED b's	.9227	-.0058
WEIGHTED b's	.9186	.0029
TRUE SCORE	.8902	.0410
ICC (H)	.9077	.0265
ICC (S/L)	.9071	.0265
MLE	.9071	.0297

See key on p.62 for identification of techniques.

CHAPTER 5

DISCUSSION AND IMPLICATIONS FOR THE SELECTION OF A TECHNIQUE TO TRANSFORM PARAMETERS TO A COMMON METRIC IN ITEM RESPONSE THEORY

The results of Chapters 3 and 4 suggest that a relatively cautious approach should be taken by researchers when chosing a technique for placing two sets of independently estimated parameters on a common metric. The findings of the simulation section of this study (Chapter 3) did not indicate that one technique possesses a uniform advantage over other techniques across the various conditions examined. The results suggest that the choice of equating technique by the researcher may be influenced by such factors as sample size, test length, and differences between the two distributions of ability. The findings of the analyses involving real data (Chapter 4) suggest that under some circumstances, the transformation of scale estimated by even the simplest techniques may be satisfactory.

Before any specific recommendations regarding the appropriate choice of equating technique are made, the results of Chapters 3 and 4 will be reviewed and discussed in detail. The recommendations made later in this chapter will be based on these observations and insights.

Discussion of Simulation Results

Comment on Experimental Design

One major limitation of the design of the simulation study involves the effects of sampling error of the equating constants on the RMSE criterion. If the analyses for a particular condition were replicated, it is possible that a different rank ordering of the techniques would be observed.

Each estimated transformation involves the estimation of two parameters, the A and the B equating constants. For a given technique, under a specified condition, each of these parameters possesses a specific standard error of estimate. The smaller the standard error of estimate, the closer we would expect the estimated transformation to be to the true transformation over numerous replications of the experiment (assuming all the techniques are unbiased). If the standard error of estimate for a technique were relatively large, we would expect to observe a wide range of estimated scale transformations over numerous replications, some close to the true transformation and others very far from the true transformation. Notice for techniques with relatively large standard errors of estimate, it is difficult to predict how close the estimated transformation will be to the true transformation on any one replication of the experiment. Predictions are usually made in terms of expected differences over many replications of the experiment. Techniques which possess smaller standard errors of estimate are desired over techniques which possess larger standard errors of measurement. Over many replications of the experiment the technique with the smaller

sampling variance would produce estimates that are closer to the true values than those produced by the technique with the larger sampling variance.

The current simulation design can be thought of as an experiment which assesses the performance of a technique using a single observation in each of seven conditions. We would like to identify the techniques with the smallest sampling errors. For a given observation, we would expect the estimates from a technique with a small sampling variance to be close to the true values (but we may occasionally observe values that are far from the true values). Notice also that it is possible for a technique with a large error of estimate to produce values close to the true values. Thus by observing a single observation it is difficult to make accurate inferences about the standard errors of estimate for the equating techniques.

One solution to the above problem would be to perform numerious replications of the experiment under each of the seven conditions. The empirical distributions of the parameter estimates could then be used to make inferences about their true sampling distributions. If a large number of replications were performed, reliable and consistent differences in the standard errors of estimate may be detected. Unfortunately the cost of replicating all the analyses in each condition is prohibitive with the resources currently available.

To gain some insight into making inferences about the performance of the techniques using a single observation, the analyses for Condition 2 were replicated an additional four times for a total of five replications. Each replication involved the sampling of new samples of

ability parameters, the generation of dichotomous responses, estimation of item and person parameters, and finally the estimation of the equating constants. The analyses are summarized in Tables 16 and 17. Table 16 displays the RMSE values and their rank orders for each technique for each of the five replications. The RMSE values are in parentheses, with the rank order of the RMSE across techniques for a given replication displayed directly to the right. Table 17 displays the mean, median, and range of RMSE values for each of the seven techniques for the same replications.

Notice (from Table 16), as anticipated the rank ordering of techniques does not remain constant across replications of the experiment. As discussed earlier this result is most likely due to sampling error of the equating constants. Table 17 gives some insight into the effect of sampling error on the RMSE criterion for each technique. Notice that the range of RMSE values for the three b parameter techniques is relatively large compared to the remaining techniques.

The results of Tables 16 and 17 indicate that small differences in RMSE values between two techniques should not be interpreted as evidence for differential performance. It is likely that small differences may be due to sampling error. Only large discrepancies in RMSE values should be treated as significant differences in performance.

The reader is cautioned against directly applying the results displayed in Tables 16 and 17 to the interpretation of results in the remaining six conditions. First, these results are based on a small number of replications and are also subject to the effects of sampling

Table 16

RMSE Values and Rank Orders
for Condition 2
for Five Replications of the Analysis

	Replication				
TECHNIQUE	1	2	3	4	5
ALL b's	(.692)7	(.390)7	(.715)7	(.713)7	(.864)7
SELECTED b's	(.050)2	(.163)6	(.125)5	(.167)5	(.070)4
WEIGHTED b's	(.076)4	(.096)4	(.198)6	(.223)6	(.118)6
TRUE SCORE	(.087)6	(.075)1	(.065)1	(.082)3	(.054)2
ICC (H)	(.047)1	(.093)3	(.123)4	(.079)2	(.069)3
ICC (S/L)	(.081)5	(.090)2	(.089)2	(.076)1	(.034)1
MLE	(.055)3	(.109)5	(.115)3	(.108)4	(.076)5

Table 17

Summary of Five Replications
of Condition 2

Technique	Mean	Median	Min-----Max	Range
ALL b's	.6748	.7133	.3904--.8643	.4738
SELECTED b's	.1151	.1255	.0501--.1673	.1172
WEIGHTED b's	.1421	.1178	.0757--.2229	.1472
TRUE SCORE	.0723	.0746	.0536--.0869	.0334
ICC (H)	.0821	.0789	.0471--.1228	.0757
ICC (S/L)	.0738	.0809	.0336--.0899	.0563
MLE	.0925	.1077	.0551--.1147	.0597

error. Second, the sampling error of a given technique might be expected to vary from condition to condition. Standard errors of estimate in general would be expected to increase as sample sizes and test length decrease, and as overlap of the two ability distributions decreases. The main point to be stressed is that differences in RMSE values should not be translated into literal differences in performance, but rather should be interpreted in the context of sampling error of the equating constants.

Summary of Simulation Results

Table 18 summarizes the results of the simulation study presented in Chapter 3 (Tables 6 through 12). In parentheses are the RMSE values for each technique, for each condition, copied from the last column of Tables 6 through 12. To the left of each RMSE value is the rank order (from smallest to largest) of the seven RMSE values for a given equating technique, rank ordered across conditions. Directly to the right of each RMSE value is the rank order (also from smallest to largest) of the RMSE values for a given condition, rank ordered across techniques.

Below, the rank ordering of RMSE values for a given equating technique are examined. Each technique is examined in turn. Some theoretical predictions and considerations from Chapter 2 are integrated with the empirical findings of Table 18.

Table 19 lists three sets of comparisons that are of special interest: Test Length (Conditions 1,6,7), Sample Size (Conditions 1,2,3), and Ability Distribution Overlap (Conditions 4,2,5). As can be

Table 18

Root Mean Squared Error Values

and Rank Orders for

Simulation Conditions

	Condition						
TECHNIQUE	1	2	3	4	5	6	7
ALL b's	2(.44)7	4(.69)7	5(1.04)7	6(1.49)7	7(1.64)7	3(.69)7	1(.27)6
SELECTED b's	5(.123)6	2(.050)2	4(.096)5	1(.020)1	6(.189)5	3(.083)5	7(.213)5
WEIGHTED b's	3(.074)5	4(.076)4	5(.116)6	1(.053)5	6(.201)6	2(.067)1	7(.278)7
TRUE SCORE	3(.054)4	5(.087)6	2(.033)1	1(.022)2	6(.113)1	7(.123)6	4(.079)1
ICC (H)	2(.039)2	3(.047)1	4(.064)2	1(.025)4	6(.137)3	5(.072)2	7(.149)3
ICC (S/L)	2(.051)3	4(.081)5	5(.082)3	1(.024)3	7(.126)2	3(.080)4	6(.121)2
MLE	1(.024)1	2(.055)3	5(.087)4	3(.056)6	6(.153)4	4(.079)3	7(.194)4

	Summary of Conditions						
N	1000	500	250	500	500	1000	1000
n	60	60	60	60	60	30	15
Comp Mean	-.5	-.5	-.5	0	-1	-.5	-.5

Table 19

Comparisons

Effect	Conditions		
Test Length	1(60 Items)	6(30 Items)	7(15 Items)
Sample Size	1(1000 Subs)	2(500 Subs)	3(250 Subs)
Distribution	4(x=0)	2(x=-0.5)	5(x=-1.0)
	Small ⟶ (Expected RMSE) ⟶ Large		

seen from Table 4, within each set of comparisons listed in Table 19, only one of the three factors (Sample size, Test Length or Distribution Overlap) varies, while the values of the other two factors remain constant. By examination of these sets of conditions, some insight into the effect of each of these factors on each technique may be obtained. Notice however that sampling error of the equating constants may in large part influence the observed ordering of the RMSE values for each of these contrasts. If for a particular technique the expected ordering of RMSE values is not observed, we should not conclude that the factor of interest has no effect, but rather should interpret the ordering of RMSE values in the context of sampling error.

1. b-parameter Equating (using all the $\underline{b}i$'s). This technique finds the transformation of the comparison group scale that equates the first two moments of the two distributions of estimated b-parameters.

The sample size comparison (Conditions 1, 2 and 3) shows the expected ordering. This rank ordering of RMSE values suggests that as sample size decreases (from 1000 to 500 to 250) our RMSE values increase (from .44 to .69 to 1.04). This finding may have been anticipated on theoretical grounds, for as sample size decreases the standard error of our b-parameters increases. This increase in the standard error of the b values from conditions 1, 2 to 3 appears likely to have produced the observed ordering of RMSE values for those conditions. The other two comparisons of interest, Test Length (Conditions 1, 6 and 7) and overlap of ability distributions (Conditions 4, 2 and 5) did not display the expected rank orderings.

Perhaps the most important observation to be made concerning the RMSE values for the first technique is to note their large values relative to the other values in the table. None of the transformations estimated by the simple b-parameter technique appears close to the true transformation. Thus, for a test with item parameters similar to those specified in Table 3 and under conditions similar to those examined, the simple b-parameter technique which incorporates all the values to estimate the transformation appears unsatisfactory.

2. b-parameter Equating (using well estimated $\underline{b}i$'s). The objective of this technique was to obtain a smaller set of better estimated $\underline{b}i$'s from which to compute the transformation of scale. This was accomplished by excluding items with extreme estimated difficulty values or small estimated discrimination values.

Table 18 reveals a substantial reduction across each of the seven conditions, of the RMSE values for the present technique. That is, by removing the items with large standard errors from the items used to estimate the sample moments, a substantial improvement of the estimated transformtion is observed.

Of the three comparisons that examine the effects of Test Length, Sample Size, and Ability Distribution Overlap, only the latter (Ability Distribution Overlap) displays the anticipated rank order of RMSE values. Notice however that the test length comparison is partially confounded by the exclusion of some items in each condition (due to large standard errors). Table 20 lists the actual number of items selected under each condition used to compute the transformation of

Table 20

Number of Items Selected by
Technique 2 for
Estimation of Sample Moments

Condition	Number Selected	Total no. of Items
1	42	60
2	43	60
3	42	60
4	45	60
5	40	60
6	17	30
7	10	15

scale. These are the number items determined to possess relatively small standard errors by examination of the estimated a and b parameters. Rather than basing our transformation on the full 60, 30 and 15 items for Conditions 1, 6 and 7; the transformation is actually based on 42, 17 and 10 items respectively.

Although this b-parameter technique did very well in Conditions 2 and 4, it is important to note that even after eliminating items with large standard errors, that there is no guarantee that the transformation based on the remaining items is adequate under any of the conditions examined.

3. b-parameter Equating (using weighted bi's). This technique controls for the effects of poorly estimated bi's by the use of weights that are inversely proportional to the estimated standard errors of the estimated item difficulties.

With respect to the first technique, there is a substantial reduction across most conditions, of the RMSE values. However, it is surprising to see that this technique did not display a large improvement over the second technique for most of the Conditions in Table 18. Additional evidence for this lack of large improvement is displayed in Table 16. Notice that for the replications of Condition 2, this technique displayed larger RMSE values than the restricted b technique in four of the five replications.

One hypothesis consistent with the above results deals with the estimator of the standard error of the b parameters (see Chapter 2 for a review). The formulas used for estimating the standard error are

asymptotic in nature, and converge to the true values as N (sample size) tends towards infinity. For relatively small samples (say 500 or less) the estimated standard error may be a poor approximation to the true standard error. By examining Table 18, we can observe that the two conditions in which the present technique possesses smaller RMSE values than the second technique (Conditions 1 and 6) are conditions which possess samples of 1000. Although this reversal of ordering for these two conditions may be atributed to sampling error, the results are consistent with the hypothesis that sample sizes of 500 or less are too small to produce accurate estimates of the standard errors of the difficulty parameters.

Both the Sample Size and Ability Distribution Overlap comparisons displayed the expected ordering of RMSE values. The Test Length comparison did not display the anticipated ordering.

4. True Score Equating (Stocking & Lord). This technique finds the linear transformation necessary to develop a common metric by using estimated true scores.

For five of the seven conditions, the RMSE values for the True score technique are smaller than the corresponding RMSE values of the three b parameter techniques. Thus, over most of the conditions examined, the scale transformation estimated by the True score technique appeared closer to the true transformation than those estimated by any of the three b-parameter techniques. Additional evidence for the superior performance of the True score technique is displayed in Table 16. The True score technique displayed smaller RMSE values than all three

b-parameter techniques in four of the five replications of Condition 2.

Of the three comparisons that examine the effects of Test Length, Sample Size, and Ability Distribution Overlap, only the latter (Ability Distribution Overlap) displayed the anticipated rank ordering of RMSE values.

5. ICC equating (Haebara) and
6. ICC equating (Segall & Levine)

Both these techniques estimate the scale transformation by examining the sum (across items) of the weighted sum of squared differences between corresponding ICC's. These two techniques differ with respect to the way that the squared differences for each ICC are weighted. The weighting scheme suggested by Haebara (1980) weights those segments of the estimated ICC differences in accordance with the relative frequency of examinees falling in the region. The weighting scheme suggested by Segall & Levine (1983), however, is formed from the product of the two empirical pdf's from each group. This weight function is largest over the range where the overlap of the two estimated distributions of ability is the greatest, and zero where there is no overlap. The goal here, remember was to place the heaviest emphasis on that portion of the squared difference between corresponding ICC's that is relatively well estimated in both groups.

Both sum of squares techniques out performed the three b-parameter techniques in almost all the conditions examined (see Table 18). The two sums of squares techniques produced RMSE values that were smaller than those produced ty the True score method in 3 out of the seven

conditions examined. From Table 18 we can observe that Haebara's method produced relatively smaller RMSE values than the Segall & Levine method in all but two of the seven conditions examined. From Table 16 however a different ordering of RMSE values is displayed for the two techniques. The Segall & Levine method displays smaller RMSE values than Haebara's method in four of the five replications of Condition 2. This reversal of RMSE ordering for these two techniques is most likely due to sampling error. Although there may be real differences in performance between the two methods, these differences, if they exist, appear too small to be detected by the present design.

A closer examination of some intermediate results suggest one alteration to the Segall & Levine method that may improve its performance. The weight function suggested by Segall & Levine is formed by taking the square root of the product of the two empirical pdf's (after transforming the comparison group distribution to the base group metric). This weighting scheme resulted in a weight function that contains a relatively large number of zero elements. As a result the weighted sums of squared difference between corresponding ICC's was formed from a relatively small number of points because of the large number of zero elements contained in the weight function. The sum of squared differences estimated from a small number of points probably resulted in a less accurate estimate of the sums of squares criterion than if the squared differences had been evaluated at a larger number of points. One improvement to this method would involve recomputing the weight function in such a manner that it examined only the range of distribution overlap, thus avoiding the problem associated with zero

elements.

In Chapter 2 it was predicted that when the two distributions of ability were roughly equal, the performance of the two sum of squares techniques should produce similar results. This prediction is confirmed by examining the RMSE values of Condition 4 for the two techniques.

7. Equating Based on Vectors of Item Parameter Differences. This technique finds the scale transformation that maximizes an approximation to the the likelihood of observing the vectors of item parameter differences. In its present form, this technique relies on several approximate properties of maximum likelihood estimators: (1) maximum likelihood estimators are approximately normally distributed with mean equal to the true parameter value; and (2) the asymptotic variance covariance matrix for these estimates may be obtained from the inverse of an approximated information matrix. The variance covariance estimates are used to define the objective function that is used to estimate the equating constants.

Over most of the conditions and replications (Tables 16 and 18) examined the RMSE values for the MLE technique appear larger than the corresponding values for the True score and sums of squares techniques. One possibility is that this observed ordering is due to sampling error, as discussed earlier. These results are also consistent with the hypothesis that the observed performance may be explained by the heavy reliance of this technique on the asymptotic propertices of maximum likelihood estimators. Some support for this explaination is achieved by examination of Condition 1, (where there are 1000 examinees and 60

items). In this condition, with a relatively large number of subjects, the MLE procedure displayed smaller RMSE values than in other conditions with smaller samples and shorter tests. It also performed relatively well compared to the other methods (although this result may not be reliable). It may be that sample sizes of 500 and test lengths of 30 are insufficient to yeild good estimates of covariances matrices and normally distributed estimates. Further analyses however would be needed to confirm this hypothesis.

Of the three comparisons that examine the effects of Test Length, Sample Size, and Ability Distribution Overlap, only the latter (Ability Distribution Overlap) failed to display the anticipated rank ordering of RMSE values.

Discussion of Real Data Results

As discussed in Chapter 4, Study I appears to indicate that none of the techniques were adversely effected by violations to the IRT model. This is evidenced by the close agreement of all the estimated equating constants to their expected values. Similar results were obtained from Study II. Although the estimated transformations were not as close to the expected values as were those from Study I, the estimated equating constants from Study II all appear in fairly close agreement.

One of the most surprising and important findings of Chapter 4 was the excellent performance of the simple b parameter equating technique. Notice that these results contradict the findings of the simulation

results (Chapter 3), which showed this technique produced poor estimates under all the conditions examined. It may be possible, however, to reconcile these findings by considering differences in the heterogeniety of the samples used to estimate the item parameters, as well as the relative difficulty of the items with respect to these samples.

Remember that b parameter values and the ability parameters "θ" are measured on the same scale. If the b value for a particular item possesses a value close to many of the true ability parameters, the estimate of that b value will possess a small standard error. If on the other hand, the b value for a particular item possesses an extreme value, far from most ability parameters, the estimate of that b value will possess a relatively large standard error.

One hypothesis for the discrepant findings between the simulation and real data studies is that there exists a different relation between the difficulty parameters and ability distributions of the two analyses. It may be that the test used to generate the simulated item responses contained many more items with extreme b values than did the real test examined in Chapter 4. These extreme b values, for the simulation analyses, resulted in poor estimates of the scale transformation. This hypothesis is examined in further detail in the following section.

The appropriateness of the simple b parameter technique is an especially important issue. It is probably the most used technique for transforming parameters to a common metric and thus deserves special attention. One the the most widely used estimation programs LOGIST (Wood, et al.) allows the metric of the theta scale to be specified by standardizing on the estimated b values. That is, there is an option

which sets the unit and origin of the theta scale to values that result in the estimated item difficulties possessing a mean of zero and a standard deviation of one. Notice that if two sets of parameters are estimated independently, this option would automatically equate the first two moments of the distribution of estimated item difficulties. This procedure is identical to the simple b parameter technique discussed in this paper. Notice that a simple rule that examined the estimated parameter values from LOGIST to judge the appropriateness of the simple b parameter technique would be extremely useful. One such technique is developed in the following section. As a basis this rule uses the results of the simulation and real data analyses of Chapters 3 and 4.

Recommendations for the Selection of Appropriate Techniques for Transforming Parameters to a Common Metric

The first issue to be addressed in this section is the specification of a simple rule for governing the use of the simple b parameter technique. The discrepant findings of Chapters 3 and 4 will be reviewed in detail and used as a basis for developing this criterion.

The second goal of this section is the specification of general guidelines concerning the use of all the equating techniques with respect to certain test and sample characteristics. These guidelines will also incorporate the results of Chapters 3 and 4, and are discussed

in the latter portion of this section.

Guidelines for Use of the Simple b-Parameter Technique

The results of Chapter 3 indicate that the simple b parameter technique performed poorly under all seven conditions examined. The results of Chapter 4, the real data analyses, indicated satisfactory performance by the equating procedure. The most likely explaination for the difference in performance can be traced to the differences in the distributions of the difficulty parameters, relative to the ability distributions. That is, for the simulation (Chapter 3) conditions, the b values were specified in a manner that resulted in a large number of extreme values (relative to the distribution of ability examined). These extreme values, of course, possess large sampling errors, which in turn result in a poor estimated scale transformation. The MAT, on the other hand, appears to have the heaviest concentration of b values in the region with the heaviest concentration of thetas. This b parameter - ability distribution relationship produces well estimated difficulty values, which in turn result in a well estimated scale transformation.

To add credibility to the above explaination, Figures 7 through 12 display the relation between the distribution of difficulty parameters and the distribution of ability, for the MAT and simulation studies. In each figure, the S-shaped curve (extending from the lower left hand corner to the upper right hand corner) represents the cumulative distribution function of the ability parameters. Each b value is represented by a pair of horizontal and vertical lines, superimposed on the same figure. Thus each b parameter for an item is represented by

one vertical and one horizontal line. The vertical line, terminating on the theta axis, represents the value of the parameter. The horizontal line for the same item, terminating on the cdf axis, indicates the proportion of thetas falling at or below the corresponding b value.

Figures 7 and 8 display the relation between the estimated b values and ability parameters for the MAT (Chapter 4, Study I). From either of the base or comparison group calibrations, we observe a heavy concentration of b values in the range -1.7 to +1.7 (vertical lines). From examination of the horizontal lines for those same items, we observe very few items that fall in the extreme tails of the cdf. Thus, almost all the b values fall in a range surrounded by a substantial number of thetas.

Figures 9 and 10 display the relation between the estimated b values and ability parameters for the shorter 45 item version of the MAT (Chapter 4, Study II). Again from either the base or comparison group calibrations, we observe a heavy concentration of b values in the range -1.7 to +1.7. As before, we observe practically no items falling in the extreme tails of the cdf. Thus, here also the b values fall in a range at or near a substantial number of examinees.

Figures 11 and 12 however, suggest an entirely different relation. These figures display the relation between the true b parameter values used to generate the simulated responses (Chapter 3) and the true ability parameters for 1000 subjects (for Condition 1). Remember that these b values were sampled from a uniform distribution in the range -3 to +3. The uniformity of these values is evidenced by the roughly even scatter of the vertical lines. From examination of the horizontal

Figure 7

Relation of Distribution of Estimated Difficulty Parameters with Cumulative Distribution of Estimated Person Parameters

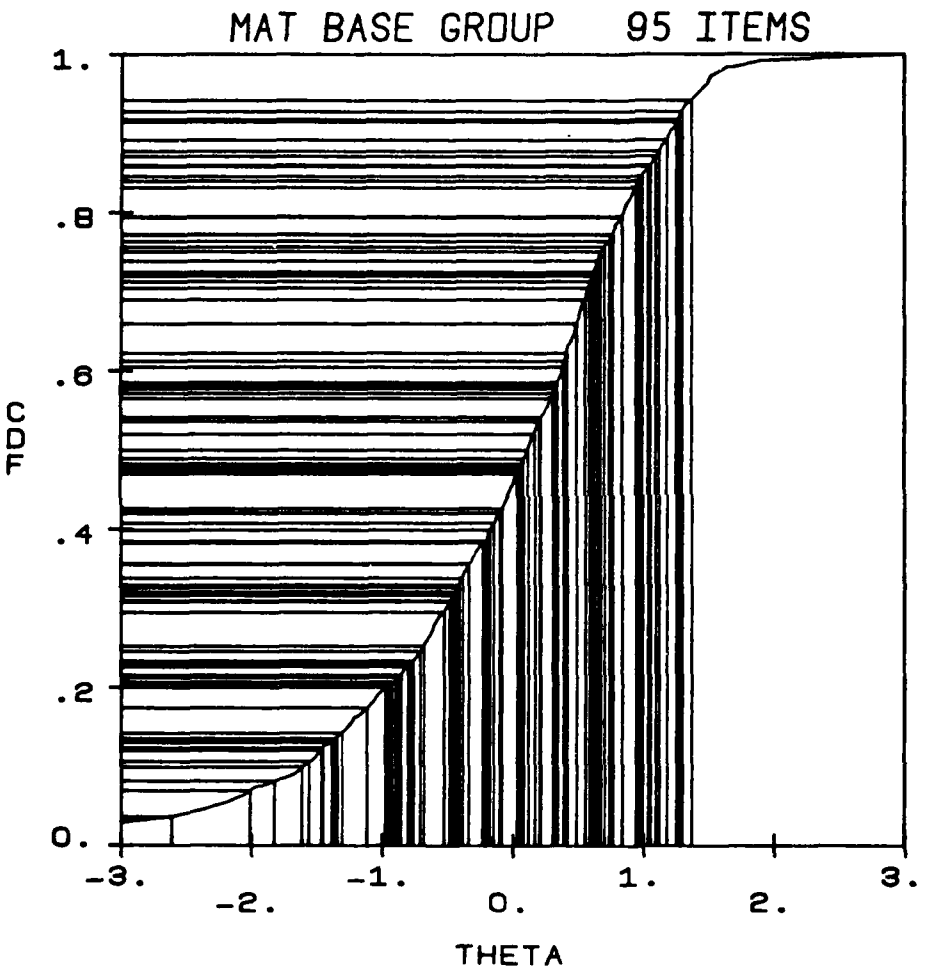


Figure 8

Relation of Distribution of Estimated Difficulty Parameters
with Cumulative Distribution of Estimated Person Parameters

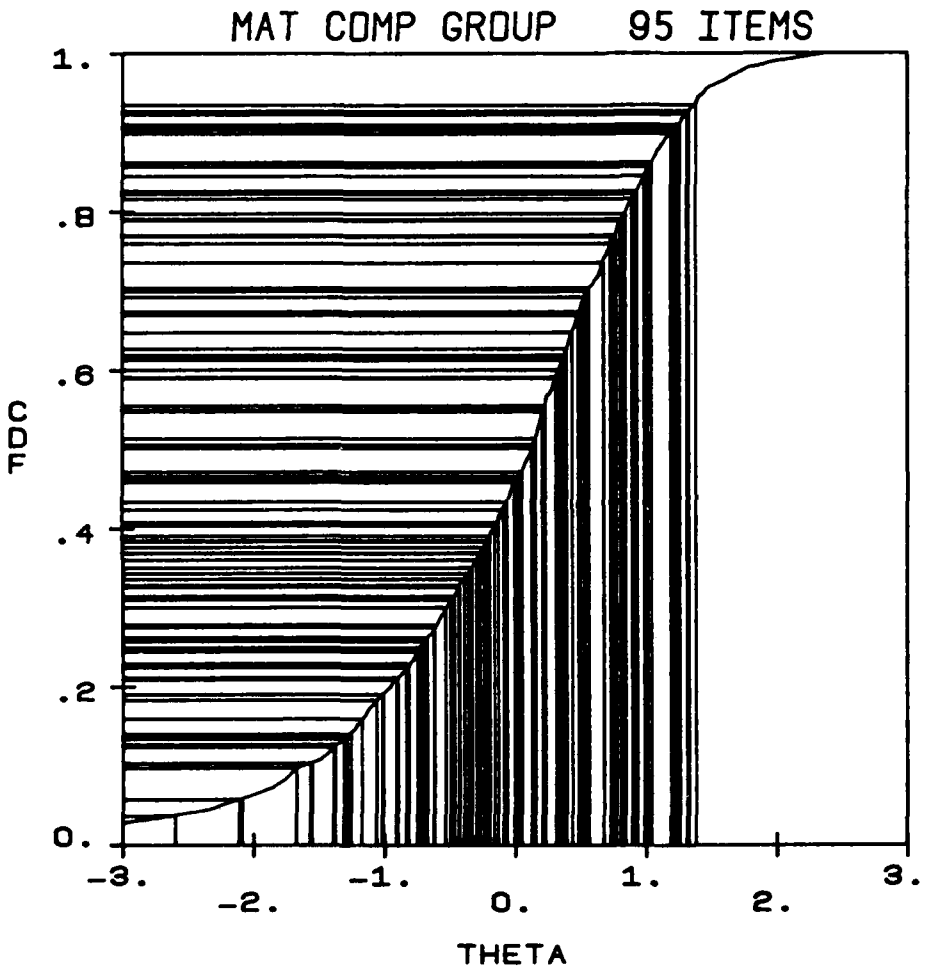


Figure 9

Relation of Distribution of Estimated Difficulty Parameters
with Cumulative Distribution of Estimated Person Parameters

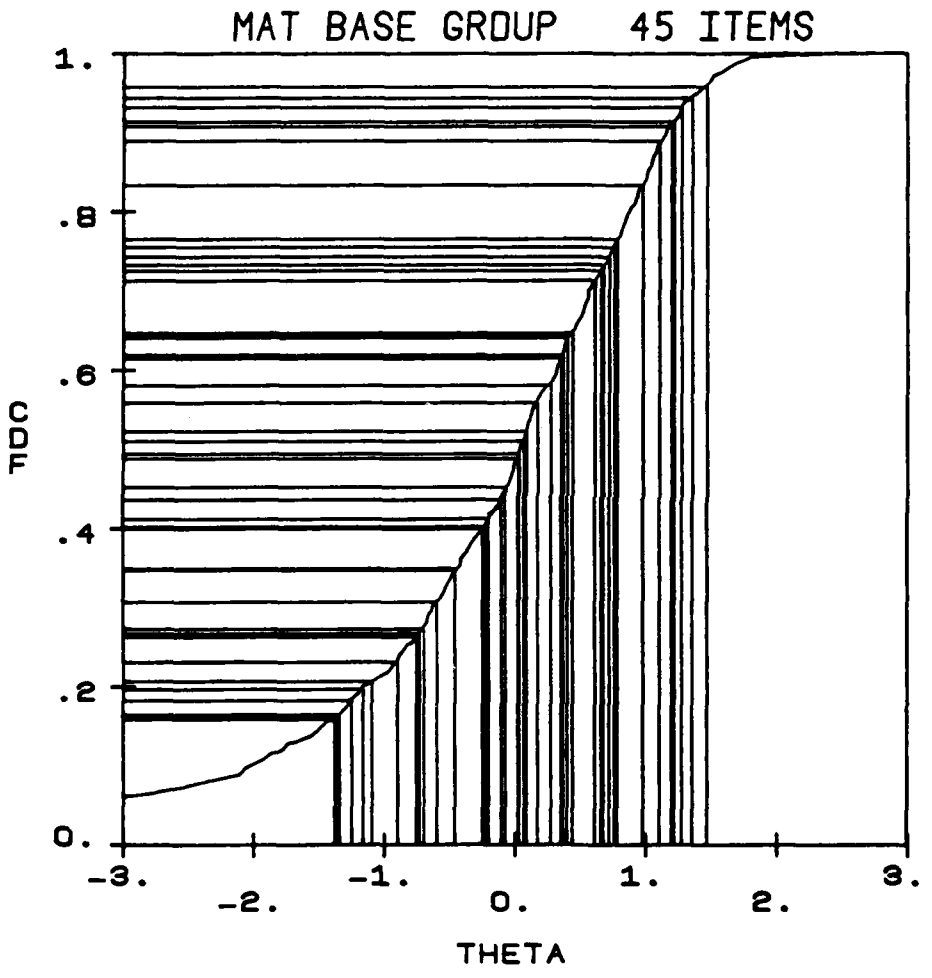


Figure 10

Relation of Distribution of Estimated Difficulty Parameters with Cumulative Distribution of Estimated Person Parameters

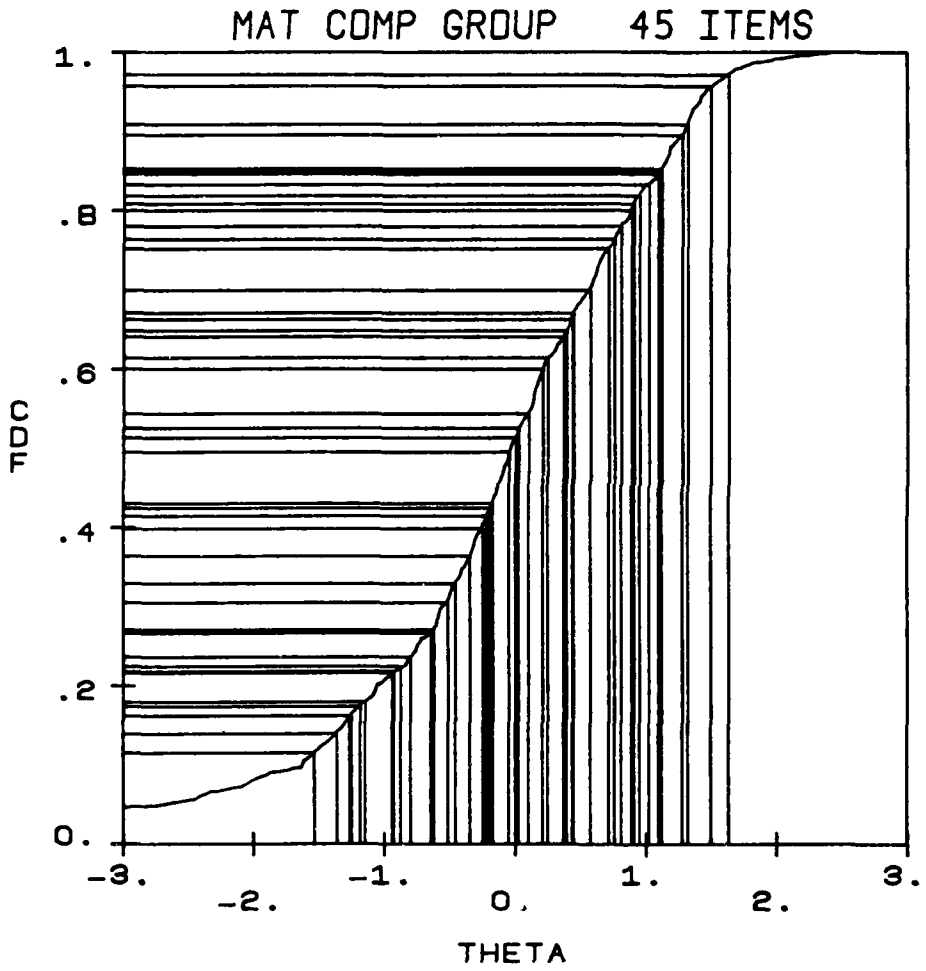


Figure 11

Relation of Distribution of True Difficulty Parameters used in the Simulation Analyses with the Cumulative Distribution of True Person Parameters (from Condition 1)

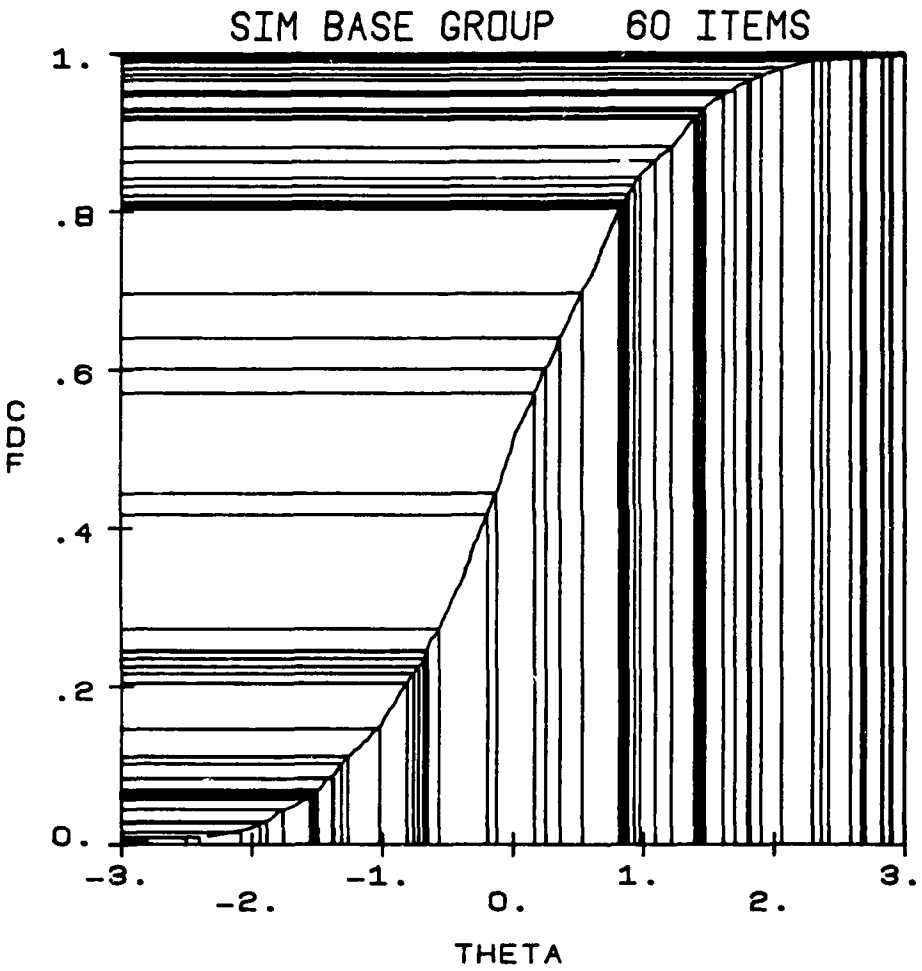
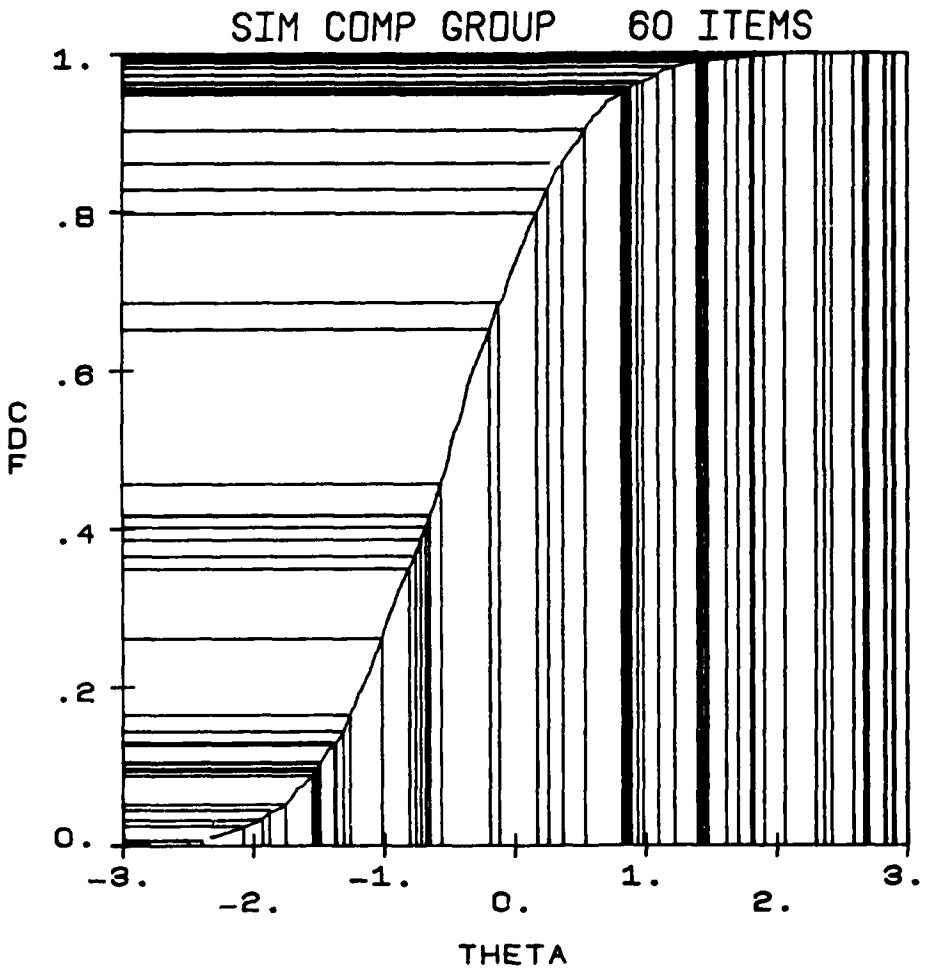


Figure 12

Relation of Distribution of True Difficulty Parameters used in the Simulation Analyses with the Cumulative Distribution of True Person Parameters (from Condition 1)



lines, for these same items, we observe a very sparse concentration of values near the center of the cdf, and very heavy concentrations near the extremes of the cdf.

Figures 11 and 12 display very dramatically the differences in the distribution of difficulty values between the real and simulation portions of this study. In the simulation portion of the study, there were many b values concentrated at the extreme tails of the cdf. These b values fall in a range surrounded by few or no examinees, thus possessing large sampling errors. These large sampling errors resulted in poor estimated transformations from the simple b parameter equating technique.

It is interesting to note that had the b values for the simulation parameters been sampled from a normal, rather than uniform distribution, results similar to those found in Figures 7 through 10 for the MAT, would probably have been observed.

Table 21 summarizes some key information found in Figures 7 through 12. For each of these figures, the number of b values falling in the extremes of the cdf were tabulated. Here, an extreme b value is one in which fewer than 5% of the ability parameters possess more extreme values. Thus, any b values falling in the cdf regions 0 to .05, and .95 to 1., were considered extreme as listed in Table 21. These are the b values likely to possess the largest standard errors.

As can be observed from Table 21, the percent of simulation b values falling in the extreme cdf range (40% and 58%) is much higher than any of the MAT calibrations (ranging from 1% to 4%). This explains very satisfactorily the differences in performance of the simple b parameter

Table 21

Frequency and Percent of b Values
with Extreme Associated cdf Values

Test	Test Length	Sample Size	Frequency		Percent	
			Base	Comp	Base	Comp
MAT	95	1000	1	1	1%	1%
MAT	45	500	1	2	2%	4%
Simulation Condition 1	60	1000	24	35	40%	58%

equating technique in the simulation and real data analyses. In the simulation analyses, a large proportion of the b values possessed large standard errors, while in the real data analyses only 1 to 4 percent at most, possessed relatively large standard errors.

Notice that Table 21 suggests a relatively useful criterion. When 4% or less of the b values fall in the extreme cdf range (as defined above) for each group, and none of these b values possess a corresponding cdf value of 0 or 1, the simple b parameter technique appears to produce satisfactory results. Thus a criterion of 4 to 5 percent (with none of the items having cdf values of 0 or 1), may be very useful in determining the adequacy of the simple b parameter equating technique. Notice however, that a cut of 4% may represent a relatively conservative estimate of the percent of items allowed to possess large values relative to the distribution of ability. Further study may show that slightly larger percentages are admissible.

Notice, also that one nice feature of the above guideline, is that it is metric free, and can be used on any set of parameter estimates, no matter how the unit and origin were specified. Thus, it would be possible, for example, to estimate simultaneously the item parameters and standardize on the estimated b parameters for the two groups independently. Then, the above procedure could be used on each set of estimated parameters to check the adequacy of this standardization, after the standardization had already been performed.

Recommendations for Appropriate Use of the Seven Equating Techniques

As mentioned at the beginning of this chapter, a relatively cautious

approach should be taken by the researcher when chosing a technique for placing two sets of independently estimated parameters on a common metric. Under the best of circumstances all the techniques appear to perform well and transforming the parameters can be performed using one of the simple b parameter techniques. Under the worst of circumstances, the more complicated sums of squares and MLE techniques offer a clear advantage over the simple b parameter techniques, and their use is encouraged.

The first step, should be one of determining how well the simple b parameter technique is suited to the test and ability distribution at hand. The analysis presented in the previous section may provide very useful insights into the suitability of the simple b parameter technique. Remember, for each group separately, the estimated ability parameters are sorted, and the proportion of thetas at or below each b value is computed. If there is a large number of the b values with extreme proportions, for either group, the researcher should consider using one of the other techniques. If there are a small number of b values in each group with extreme values, all the techniques would be expected to perform relatively well. Thus, in the latter instance, choice of equating method is not critical.

When a relatively large number of b values with extreme proportions have been encountered, choice of equating technique may be further influenced by such factors as sample size and test length. Again for relatively large samples (1000 or more in each group) and relatively long tests (60 or more items) the True Score, Sums of Squares, and MLE procedures would be expected to produce satisfactory results.

For shorter tests and smaller samples, the True Score or either of the ICC techniques would be likely to produce the most favorable results. Because of the heavy reliance of the MLE technique on certain asymptotic properties, its use is not recommended with small samples and short tests.

Recommendations for Further Study

The results and insights gained from the current research raise a number of issues deserving further investigation. Several of these areas are described below.

Asymptotic Sampling Variance of Item Parameters

The results of portions of the simulation analyses raise several questions concerning the relation amoung the asymptotic properties of maximum likelihood estimates of item parameters, the estimators of the standard error of these parameters, and sample size. The results concerning the performance of the weighted b parameter technique appear consistent with the hypothesis that samples of 1000 may be needed before the estimates of the standard error of the b parameters are close approximations to the true values. The results concerning the performance of the MLE equating technique also appear consistent with the hypothesis that samples of 1000 may be needed before the asymptotic properties of the item parameters, and there variance covariance estimates are realized. Further investigation into the relation between

sample size and the assumed asymptotic properties of the item parameters for the logistic model would be relevant to many other areas of IRT as well as the current research.

Guidelines for Simple b-Parameter Technique

Remember from the previous section that 4% was suggested as the maximum proportion of items in a test possessing extreme cdf values for the acceptable use of the simple b parameter technique. Remember that this criterion was selected on a basis of the results from the actual MAT analyses, where the b parameter technique performed relatively well. As indicated earlier, a criterion of 4% of the test items may represent a relatively concervative criterion. A systematic study into the effects of varying the number of items in a test with extreme b values may indicate whether in fact a criterion of 4% is too concervative.

Improvements to Equating Techniques

Results from the analyses produced several insights for further modifications to several of the techniques.

True Score Equating Technique (Stocking & Lord). The criterion minimized by this technique only involves minimizing the sums of squared differences between true scores for one of the two groups of examinees. An increase in power may perhaps be achieved by adding another term to the loss function which reflects the analogous term for members of the other group of examinees, thus incorporating transformed and untransformed true score estimates from both samples.

Sums of Squares Equating Technique (Segall & Levine). The weight

function used in calculating the criterion contains zero's over a substantial range of theta, producing a weighted sums of squares estimate for each item that was formed on the basis of a very small number of points. One improvement to this technique would involve a modification to the weighting scheme that evaluated the ICC using a larger number of points in the same range of ability distribution overlap. This would produce a more accurate estimate of the weighted sums of squares criterion, which in turn may improve the performance of the technique.

APPENDIX

Fortran Computer Programs Used to Estimate Scale Transformation for All Seven Equating Techniques

EQUATE - Subroutine

This routine estimates the linear transformation necessary to place two independently estimated sets of parameters on a common metric. Seven approaches are available. These are discribed in Chapter 2. This routine makes several calls to to IMSL (Version 9) subroutines. All other subroutines are listed below. Program was written for use on CDC system using Fortran Extended Version 4.

```
EQUATE(PARB,PARC,IROW,NITEMS,THETAB,THETAC,NSUBSB,
       NSUBSC,NPAR,IOPT,A,B,IOUT,TIME,IERROR)
```

PARB: Matrix of NITEMS rows by 3 columns containing the estimated item parameters for the base group. Row 1 contains the item parameters for item 1, Row 2 contains the item parameters of item 2, etc. Column 1 contains the a-parameters, column 2 contains the b-parameters and column 3 contains the c-parameters. If using the 2 parameter model all values in column 3 should be set to zero.

PARC: Matrix of NITEMS rows by 3 columns containing the estimated item parameters for the comparison group. Format is same as PARB.

IROW: Row dimension of PARB and PARC exactly as specified in the calling program.

NITEMS: Number of items in PARB and PARC.

THETAB: Vector of length NSUBSB containing ability parameter estimates for base group examinees. A value of 999 is treated as missing.

THETAC: Vector of length NSUBSC containing ability parameter estimates for comparison group examinees. A value of 999 is treated as missing.

NSUBSB Number of subjects in base group (including 999's).

NSUBSC Number of subjects in comparison group (including 999's).

NPAR: Number of item parameters in model. If using the 2 parameter model, NPAR=2, otherwise, NPAR=3.

IOPT: Vector of length seven. To obtain equating constant estimates for technique k, set IOPT(k)=1, where:

IOPT(1) = b-parameter equating (using all b's)
IOPT(2) = b-parameter equating (using well estimated b's)
IOPT(3) = b-parameter equating (using weighted b's)
IOPT(4) = True score equating (Stocking & Lord)
IOPT(5) = ICC equating (Haebara)
IOPT(6) = ICC equating (Segall & Levine)
IOPT(7) = MLE equating based on parameter differences

A: Vector of length seven containing the A constant estimates for techniques specified in IOPT. The estimate for technique k are in A(k).

B: Vector of length seven containing the B constant estimates for techniques specified in IOPT. The estimate for technique k are in B(k).

IOUT: Tape number specified in program statement to which error messages will be written.

TIME: Vector of length seven containing the number of CPU seconds used to estimated the equating constants. The amount of time used by technique k is in TIME(k).

IERROR: Vector of length seven containing error message codes. A value of zero for the kth element signifies that satisfactory estimates were obtained by technique k.

Subroutine Listings:

```
      SUBROUTINE EQUATE(PARB,PARC,IROW,NITEMS,THETB,THETC,
     * NB,NC,NPAR,IOPT,A,B,IOUT,TIME,ERROR)
C
      REAL PARB(IROW,3),PARC(IROW,3),THETB(NB),
     * THETC(NC),A(7),B(7),THETAB(2000),THETAC(2000),
     * TIME(7)
      INTEGER IOPT(7),ERROR(7)
C
C     INITIALIZE CONSTANTS TO ZERO
```

```
      DO 11 L = 1,7
          A(L) = 0.
          B(L) = 0.
   11     CONTINUE
C
C      CHECK FOR MISSING THETAS IN BASE GROUP
      K = 1
      DO 17 J = 1,NB
          IF(THETB(J) .EQ. 999.) GOTO 17
          THETAB(K) = THETB(J)
          K = K + 1
   17     CONTINUE
      NSUBSB = K - 1
C
C      CHECK FOR MISSING THETAS IN COMP GROUP
      K = 1
      DO 19 J = 1,NC
          IF (THETC(J) .EQ. 999.) GOTO 19
          THETAC(K) = THETC(J)
          K = K + 1
   19     CONTINUE
      NSUBSC = K - 1
C
C      SELECT SPECIFIED EQUATING TECHNIQUES:
      T1 = SECOND(CP)
      IF (IOPT(1) .EQ. 1)
     * CALL BEQUAT(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
      ERROR(1) = IER
      T2 = SECOND(CP)
      TIME(1) = T2 - T1
      T1 = T2
C
      IF (IOPT(2) .EQ. 1)
     * CALL BEQUAR(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
      ERROR(2) = IER
      T2 = SECOND(CP)
      TIME(2) = T2 - T1
      T1 = T2
C
      IF (IOPT(3) .EQ. 1)
     * CALL BEQUAC(PARB,PARC,IROW,NITEMS,THETAB,THETAC,NSUBSB,
     * NSUBSC,NPAR,A,B,IOUT,IER)
      ERROR(3) = IER
      T2 = SECOND(CP)
      TIME(3) = T2 - T1
      T1 = T2
C
      IF (IOPT(4) .EQ. 1)
     * CALL TESLRD(PARB,PARC,IROW,NITEMS,THETAB,NSUBSB,A,B,IOUT,IER)
      ERROR(4) = IER
      T2 = SECOND(CP)
      TIME(4) = T2 - T1
```

```
      T1 = T2
C
      IF (IOPT(5) .EQ. 1)
     * CALL HAEBAR(PARB,PARC,IROW,NITEMS,THETAB,THETAC,NSUBSB,
     * NSUBSC,A,B,IOUT,IER)
      ERROR(5) = IER
      T2 = SECOND(CP)
      TIME(5) = T2 - T1
      T1 = T2
C
      IF (IOPT(6) .EQ. 1)
     * CALL SLSUMS(PARB,PARC,IROW,NITEMS,THETAB,THETAC,NSUBSB,
     * NSUBSC,A,B,IOUT,IER)
      ERROR(6) = IER
      T2 = SECOND(CP)
      TIME(6) = T2 - T1
      T1 = T2
C
      IF (IOPT(7) .EQ. 1)
     * CALL EQUMLE(PARB,PARC,IROW,NITEMS,THETAB,THETAC,
     * NSUBSB,NSUBSC,NPAR,A,B,IOUT,IER)
      ERROR(7) = IER
      T2 = SECOND(CP)
      TIME(7) = T2 - T1
C
      RETURN
      END
C
C ******************** UNRESTRICTED DIFFICULTY EQUATING ******************
C
      SUBROUTINE BEQUAT(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
C
      REAL PARB(IROW,3),PARC(IROW,3),A(7),B(7)
C
C      INITIALIZE VALUES
      BTOT = 0.
      CTOT = 0.
      BTOT2 = 0.
      CTOT2 = 0.
      IER = 0
C
C      COMPUTE MEANS AND SDS
      DO 30 I = 1,NITEMS
          BTOT = BTOT + PARB(I,2)
          BTOT2 = BTOT2 + PARB(I,2)**2.
          CTOT = CTOT + PARC(I,2)
          CTOT2 = CTOT2 + PARC(I,2)**2.
   30     CONTINUE
C
      SDB = (SQRT(FLOAT(NITEMS)*BTOT2 - BTOT**2.))/FLOAT(NITEMS)
      SDC = (SQRT(FLOAT(NITEMS)*CTOT2 - CTOT**2.))/FLOAT(NITEMS)
      BMEAN = BTOT / FLOAT(NITEMS)
```

```
      CMEAN = CTOT / FLOAT(NITEMS)
C
C      COMPUTE EQUATING CONSTANTS:
      A(1) = SDB/SDC
      B(1) = BMEAN - A(1) * CMEAN
C
      RETURN
      END
C
C ***************** RESTRICTED DIFFICULTY EQUATING ********************
C
      SUBROUTINE BEQUAR(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
C
      REAL PARB(IROW,3),PARC(IROW,3),A(7),B(7)
C
C      INITIALIZE VALUES
      BTOT = 0.
      CTOT = 0.
      BTOT2 = 0.
      CTOT2 = 0.
      XNITEM = NITEMS
      IER = 0
C
C      COMPUTE MEANS AND SDS
      DO 30 I = 1,NITEMS
          IF (PARB(I,1) .LT. 0.15 .OR. ABS(PARB(I,2)) .GT. 3.0
     * .OR. PARC(I,1) .LT. 0.15 .OR. ABS(PARC(I,2)) .GT. 3.0)
     # XNITEM = XNITEM - 1
          IF (PARB(I,1) .LT. 0.15 .OR. ABS(PARB(I,2)) .GT. 3.0
     * .OR. PARC(I,1) .LT. 0.15 .OR. ABS(PARC(I,2)) .GT. 3.0)
     * GOTO 30
          BTOT = BTOT + PARB(I,2)
          BTOT2 = BTOT2 + PARB(I,2)**2.
          CTOT = CTOT + PARC(I,2)
          CTOT2 = CTOT2 + PARC(I,2)**2.
   30     CONTINUE
C
      SDB = (SQRT(XNITEM*BTOT2 - BTOT**2.))/XNITEM
      SDC = (SQRT(XNITEM*CTOT2 - CTOT**2.))/XNITEM
      BMEAN = BTOT / XNITEM
      CMEAN = CTOT / XNITEM
C
C      COMPUTE EQUATING CONSTANTS:
      A(2) = SDB/SDC
      B(2) = BMEAN - A(2) * CMEAN
C
      RETURN
      END
C
C ************** WEIGHTED DIFFICULTY EQUATING **************************
C
      SUBROUTINE BEQUAC(PARB,PARC,IROW,NITEMS,THETAB,THETAC,
```

```
     * NSUBSB,NSUBSC,NPAR,A,B,IOUT,IER)
C
      REAL PARB(IROW,3),PARC(IROW,3),THETAB(NSUBSB),
     * THETAC(NSUBSC),A(7),B(7),W(200),COVB(3,3),COVC(3,3)
C
C      INITIALIZE ERROR VAR
      IER = 0
C
C      COMPUTE COVARIANCES FOR EACH ITEM FROM EACH GROUP
      DO 101 I = 1,NITEMS
C
C      COMPUTE COV FOR BASE GROUP
      APAR = PARB(I,1)
      BPAR = PARB(I,2)
      CPAR = PARB(I,3)
      CALL COV3PL(THETAB,NSUBSB,APAR,BPAR,CPAR,NPAR,3,COVB,
     * IOUT,IER)
C
C      COMPUTE COV MATRIX FOR COMP GROUP
      APAR = PARC(I,1)
      BPAR = PARC(I,2)
      CPAR = PARC(I,3)
      CALL COV3PL(THETAC,NSUBSC,APAR,BPAR,CPAR,NPAR,3,COVC,
     * IOUT,IER)
C
C      EXTRACT LARGER OF THE TWO VARIANCE ESTIMATES
      W(I) = 1./COVB(2,2)
      IF (COVC(2,2) .GT. COVB(2,2)) W(I) = 1./COVC(2,2)
  101 CONTINUE
C
C      COMPUTE WEIGHTED MEANS
      CONST = 0.
      BMEAN = 0.
      CMEAN = 0.
      SDB = 0.
      SDC = 0.
      DO 121 I = 1,NITEMS
          BMEAN = BMEAN + PARB(I,2) * W(I)
          CMEAN = CMEAN + PARC(I,2) * W(I)
          CONST = CONST + W(I)
  121     CONTINUE
      BMEAN = BMEAN / CONST
      CMEAN = CMEAN / CONST
C
C      COMPUTE WEIGHTED SD
      DO 178 I = 1,NITEMS
          SDB = SDB + ((PARB(I,2)-BMEAN)**2.) * W(I)
          SDC = SDC + ((PARC(I,2)-CMEAN)**2.) * W(I)
  178     CONTINUE
      SDB = SQRT(SDB/CONST)
      SDC = SQRT(SDC/CONST)
C
```

```
C      COMPUTE EQUATING CONSTANTS
       A(3) = SDB / SDC
       B(3) = BMEAN - A(3) * CMEAN
C
       RETURN
       END
C
C ******************** STOCKING AND LORD ********************************
C
       SUBROUTINE TESLRD(PARB,PARC,IROW,NITEMS,THETAB,NSUBSB,A,B,
      * IOUT,IER)
C
       EXTERNAL FUNLRD
       REAL PARB(IROW,3),PARC(IROW,3),THETAB(NSUBSB),
      * A(7),B(7),X(2),W(21),SPARC(200,3),PAR(1),F(2),
      * STHETA(2000),ETAB(2000)
       INTEGER SNITEM,SNSUBS
C
       COMMON SPARC,STHETA,SNITEM,SNSUBS,ETAB
C
C      INITIALIZE ERROR VAR
       IER = 0
C
C      TRANSFER STUFF INTO COMMON ARRAYS
       DO 105 I = 1,NITEMS
           DO 101 J = 1,3
               SPARC(I,J) = PARC(I,J)
 101           CONTINUE
 105       CONTINUE
C
       DO 109 J = 1,NSUBSB
           STHETA(J) = THETAB(J)
 109       CONTINUE
       SNITEM = NITEMS
       SNSUBS = NSUBSB
C
C      INITIALIZE VALUES OF EQUATING CONSTANTS
       IF (A(2) .EQ. 0.0 .AND. B(2) .EQ. 0.0)
      * CALL BEQUAR(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
       X(1) = A(2)
       X(2) = B(2)
       CALL UGETIO(3,0,IOUT)
C
C      COMPUTE ETA FROM BASE GROUP CALIBRATION
       DO 18 J = 1,NSUBSB
           THET = THETAB(J)
           ETAB(J) = 0.
           DO 29 I = 1,NITEMS
               APAR = PARB(I,1)
               BPAR = PARB(I,2)
               CPAR = PARB(I,3)
               PROB = P3PL(APAR,BPAR,CPAR,THET)
```

```
            ETAB(J) = ETAB(J) + PROB
  29        CONTINUE
  18    CONTINUE
C
C     PERFORM MINIMIZATION
      NSIG = 3
      MAXFN = 200
      N = 2
      CALL ZSPOW (FUNLRD,NSIG,N,MAXFN,PAR,X,FNORM,W,IER)
      A(4) = X(1)
      B(4) = X(2)
C
      RETURN
      END
C
      SUBROUTINE FUNLRD(X,F,N,PAR)
C
      REAL X(2),SPARC(200,3),STHETA(2000),ETAB(2000),
     * F(2),PAR(1)
      INTEGER SNSUBS,SNITEM
C
      COMMON SPARC,STHETA,SNITEM,SNSUBS,ETAB
C
C     INITIALIZE VALUES
      A = X(1)
      B = X(2)
      DIRA = 0.
      DIRB = 0.
      D = 1.702
C
C     FOR EACH SUBJECT
      DO 500 J = 1,SNSUBS
          THET = STHETA(J)
C
C     COMPUTE DERIVATIVES OF ETA (FOR COMPARISON GROUP CALIBRATION)
      ETAC = 0.
      DETAA = 0.
      DETAB = 0.
      DO 39 I = 1,SNITEM
          APAR = SPARC(I,1)
          BPAR = SPARC(I,2)
          CPAR = SPARC(I,3)
C
C     PARTIAL DERIVATIVE OF P WITH RESPECT TO A
      DPROBA=(EXP(((BPAR*A-THET+B)*APAR*D)/A)*(THET-B)*(CPAR-1.)*
     + APAR*D)/((EXP(((BPAR*A-THET+B)*APAR*D)/A)+1+)**2*A**2)
C
C     PARTIAL DERIVATIVE OF P WITH RESPECT TO B
      DPROBB=(EXP(((BPAR*A-THET+B)*APAR*D)/A)*(CPAR-1.)*APAR*D)/((EXP
     + (((BPAR*A-THET+B)*APAR*D)/A)+1+)**2*A)
C
          DETAA = DETAA + DPROBA
```

```
          DETAB = DETAB + DPROBB
C
          APAR = SPARC(I,1) / A
          BPAR = A * SPARC(I,2) + B
          CPAR = SPARC(I,3)
          PROB = P3PL(APAR,BPAR,CPAR,THET)
          ETAC = ETAC + PROB
 39       CONTINUE
C
      DIRA = DIRA + (ETAB(J)-ETAC) * DETAA
      DIRB = DIRB + (ETAB(J)-ETAC) * DETAB
 500  CONTINUE
C
      F(1) = (-2. / FLOAT(SNSUBS)) * DIRA
      F(2) = (-2. / FLOAT(SNSUBS)) * DIRB
C
      RETURN
      END
C
C ************************* HAEBARA SUMS OF SQUARES *********************
C
      SUBROUTINE HAEBAR(PARB,PARC,IROW,NITEMS,THETAB,THETAC,NSUBSB,
     * NSUBSC,A,B,IOUT,IER)
C
      EXTERNAL HFUNCT
      REAL PARB(IROW,3),PARC(IROW,3),THETAB(NSUBSB),
     * THETAC(NSUBSC),CUTP(21),MIDP(20),HBASE(20),HCOMP(20),
     * SPARB(200,3),SPARC(200,3),X(2),H(3),G(2),W(6),A(7),B(7)
      INTEGER SNITEM
C
      COMMON HBASE,HCOMP,NINT,SPARB,SPARC,SNITEM,MIDP
C
C      INITIALIZE VARIABLES
      NCUT = 21
      NINT = NCUT - 1
      XLOWC = -3.
      XHIC = 3.
      SNITEM = NITEMS
      IER = 0
C
C      INITIALIZE CUT POINTS
      CALL ESPNT(CUTP,NCUT,XLOWC,XHIC)
C
C      INITIALIZE MIDPOINTS
      HDELT = (CUTP(2) - CUTP(1)) / 2.
      XLOWM = XLOWC + HDELT
      XHIM = XHIC - HDELT
      CALL ESPNT(MIDP,NINT,XLOWM,XHIM)
C
C      COMPUTE PROPORTIONS FOR BASE GROUP DISTRIBUTION
      CALL PEDIS(CUTP,NCUT,NINT,THETAB,NSUBSB,HBASE)
C
```

```
C     COMPUTE PROPORTIONS FOR COMPARISON GROUP DISTRIBUTION
      CALL PEDIS(CUTP,NCUT,NINT,THETAC,NSUBSC,HCOMP)
C
C     TRANSFER ITEM PARAMETERS INTO COMMON ARRAYS
      DO 29 I = 1,NITEMS
          DO 27 J = 1,3
              SPARB(I,J) = PARB(I,J)
              SPARC(I,J) = PARC(I,J)
   27         CONTINUE
   29     CONTINUE
C
C     INITIALIZE STARTING VALUES OF EQUATING CONSTANTS
      IF (A(2) .EQ. 0.0 .AND. B(2) .EQ. 0.0)
     * CALL BEQUAR(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
      X(1) = A(2)
      X(2) = B(2)
C
C     PERFORM MINIMIZATION
      CALL UGETIO(3,0,IOUT)
      N = 2
      NSIG = 3
      MAXFN = 500
      IIOPT = 2
C
      CALL ZXMIN(HFUNCT,N,NSIG,MAXFN,IIOPT,X,H,G,F,W,IER)
      A(5) = X(1)
      B(5) = X(2)
C
      RETURN
      END
C
      SUBROUTINE HFUNCT(N,X,F)
C
      INTEGER SNITEM
      REAL HBASE(20),HCOMP(20),SPARB(200,3),SPARC(200,3),MIDP(20),
     * X(2)
C
      COMMON HBASE,HCOMP,NINT,SPARB,SPARC,SNITEM,MIDP
C
C     INITIALIZE PARAMETERS
      A = X(1)
      B = X(2)
      SSC = 0.
      SSB = 0.
C
C     ACCUMULATE SUMS OF SQUARES
      DO 18 I = 1,SNITEM
          ABASE = SPARB(I,1)
          BBASE = SPARB(I,2)
          CBASE = SPARB(I,3)
          ACOM = SPARC(I,1)
          BCOM = SPARC(I,2)
```

```
          CCOM = SPARC(I,3)
C
          DO 14 J = 1,NINT
C
C              FOR COMPARISON GROUP
               TC = MIDP(J)
               TB = A * MIDP(J) + B
               SS = (P3PL(ACOM,BCOM,CCOM,TC) -
     *              P3PL(ABASE,BBASE,CBASE,TB)) ** 2.
               SSC = SSC + (SS * HCOMP(J))
C
C              FOR BASE GROUP
               TC = (MIDP(J) - B) / A
               TB = MIDP(J)
               SS = (P3PL(ABASE,BBASE,CBASE,TB) -
     *              P3PL(ACOM,BCOM,CCOM,TC)) ** 2.
               SSB = SSB + (SS * HBASE(J))
   14          CONTINUE
   18      CONTINUE
C
      F = SSC + SSB
C
      RETURN
      END
C
      SUBROUTINE PEDIS(CUTP,NCUT,NINT,THETA,NSUBS,H)
C
      REAL CUTP(NCUT),H(NINT),THETA(NSUBS)
C
C      INITIALIZE TO ZERO
      DO 10 L = 1,NINT
          H(L) = 0.
   10     CONTINUE
C
C      UPDATE FREQUENCIES
      DO 400 J = 1,NSUBS
          DO 200 K = 1,NINT
              KP1 = K + 1
              IF (THETA(J) .LE. CUTP(KP1) .AND. THETA(J) .GT. CUTP(K))
     *            H(K) = H(K) + 1.
  200         CONTINUE
  400     CONTINUE
C
C      TRANSFORM RELATIVE FREQUENCIES TO RELATIVE PROPORTIONS
      DO 500 L = 1,NINT
          H(L) = H(L) / FLOAT(NSUBS)
  500     CONTINUE
C
      RETURN
      END
C
C ********************* WEIGHTED SUMS OF SQUARES ************************
```

```
C
      SUBROUTINE SLSUMS (PARB,PARC,IROW,NITEMS,THETAB,THETAC,NSUBSB,
     * NSUBSC,A,B,IOUT,IER)
C
      EXTERNAL SLFUN
      REAL PARB(IROW,3),PARC(IROW,3),THETAB(NSUBSB),
     * THETAC(NSUBSC),CUTPC(21),MIDPC(20),HCOMP(20),SPARB(200,3),
     * X(2),W(21),A(7),B(7),F(2),SPARC(200,3),PAR(1),
     * CUTPB(21),HBASE(20),WEIGHT(20)
C
      INTEGER SNITEM
C
      COMMON MIDPC,SNITEM,SPARB,SPARC,WEIGHT
C
C     INITIALIZE VARIABLES
      NCUT = 21
      NINT = NCUT - 1
      XLOWC = -3.
      XHIC = 3.
      SNITEM = NITEMS
      CRIT = .001
      ISTAGE = 0
      IER = 0
C
C     INITIALIZE COMPARISON GROUP CUT-POINTS
      CALL ESPNT(CUTPC,NCUT,XLOWC,XHIC)
C
C     INITIALIZE COMPARISON GROUP MIDPOINTS
      HDELT = (CUTPC(2) - CUTPC(1)) / 2.
      XLOWM = XLOWC + HDELT
      XHIM = XHIC - HDELT
      CALL ESPNT(MIDPC,NINT,XLOWM,XHIM)
C
C     COMPUTE PROPORTIONS FOR COMPARISON GROUP DISTRIBUTION
      CALL PEDIS(CUTPC,NCUT,NINT,THETAC,NSUBSC,HCOMP)
C
C     TRANSFER ITEM PARAMETERS INTO COMMON ARRAYS
      DO 29 I = 1,NITEMS
          DO 27 J = 1,3
              SPARB(I,J) = PARB(I,J)
              SPARC(I,J) = PARC(I,J)
   27         CONTINUE
   29     CONTINUE
C
C     INITIALIZE STARTING VALUES OF EQUATING CONSTANTS
      IF (A(2) .EQ. 0.0 .AND. B(2) .EQ. 0.0)
     * CALL BEQUAR(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
      X(1) = A(2)
      X(2) = B(2)
C
C     FOR EACH STAGE
   88 ISTAGE = ISTAGE + 1
```

```
      AL = X(1)
      BL = X(2)
C
C     COMPUTE BASE GROUP CUT-POINTS
      DO 17 J = 1,NCUT
          CUTPB(J) = (CUTPC(J) * X(1)) + X(2)
   17     CONTINUE
C
C     COMPUTE PROPORTIONS FOR BASE GROUP DISTRIBUTION
      CALL PEDIS(CUTPB,NCUT,NINT,THETAB,NSUBSB,HBASE)
C
C     COMPUTE WEIGHT FUNCTION
      DO 52 J = 1,NINT
          WEIGHT(J) = SQRT(HBASE(J) * HCOMP(J))
   52     CONTINUE
C
C     PERFORM MINIMIZATION
      CALL UGETIO(3,0,IOUT)
C     CALL UERSET(0,LEVOLD)
      N = 2
      NSIG = 3
      MAXFN = 200
      CALL ZSPOW(SLFUN,NSIG,N,MAXFN,PAR,X,FNORM,W,IER)
C
C     CHECK FOR CONVERGENCE
      D1 = ABS(AL-X(1))
      D2 = ABS(BL-X(2))
      IF (ISTAGE .GE. 10) GOTO 603
      IF (D1 .GT. CRIT .OR. D2 .GT. CRIT) GOTO 88
C
C     FINISH
  603 CONTINUE
      A(6) = X(1)
      B(6) = X(2)
C
      RETURN
      END
C
      SUBROUTINE SLFUN(X,F,N,PAR)
C
      REAL SPARB(200,3),PAR(1),F(2),
     * MIDPC(20),X(2),
     * WEIGHT(20),SPARC(200,3)
C
      INTEGER SNITEM
C
      COMMON MIDPC,SNITEM,SPARB,SPARC,WEIGHT
C
C     INITIALIZE VALUES
      A = X(1)
      B = X(2)
C
```

```
C      COMPUTE PARTIAL DERIVATIVES
       F(1) = PDA(SPARB,SPARC,SNITEM,MIDPC,WEIGHT,A,B)
       F(2) = PDB(SPARB,SPARC,SNITEM,MIDPC,WEIGHT,A,B)
C
       RETURN
       END
C
       REAL FUNCTION PDA(PARB,PARC,NITEMS,MIDT,WEIGHT,A,B)
C
       REAL PARB(200,3),PARC(200,3),MIDT(20),WEIGHT(20)
C      PARTIAL DERIVATIVE WITH RESPECT TO A
C
C      INITIALIZE VALUES
       PDA = 0.0
       D = 1.702
C
       DO 300 I = 1,NITEMS
           ABI = PARB(I,1)
           BBI = PARB(I,2)
           CBI = PARB(I,3)
           ACI = PARC(I,1)
           BCI = PARC(I,2)
           CCI = PARC(I,3)
           DO 250 J = 1,20
               TJ = MIDT(J)
               TTJ = TJ * A + B
               WJ = WEIGHT(J)
               IF (WJ .EQ. 0.0) GOTO 250
C
C      PARTIAL DERIVATIVE OF PB WITH RESPECT TO A
       ANS=(-EXP(BBI*ABI*D+TJ*ABI*D*A+ABI*D*B)*(CBI-1.)*TJ*ABI*D)
      + /(EXP(BBI*ABI*D)+EXP(TJ*ABI*D*A+ABI*D*B))**2
C
C      COMPUTE PROBS
       PB = P3PL(ABI,BBI,CBI,TTJ)
       PC = P3PL(ACI,BCI,CCI,TJ)
C
C      COMPUTE DIRIVITIVE OF TERM
       TERM = 2. * WJ * ANS * (PB - PC)
C
       PDA = PDA + TERM
  250          CONTINUE
  300      CONTINUE
C
       RETURN
       END
C
       REAL FUNCTION PDB(PARB,PARC,NITEMS,MIDT,WEIGHT,A,B)
C
       REAL PARB(200,3),PARC(200,3),MIDT(20),WEIGHT(20)
C
C      INITIALIZE VALUES
```

```
      PDB = 0.0
      D = 1.702
C
      DO 300 I = 1,NITEMS
          ABI = PARB(I,1)
          BBI = PARB(I,2)
          CBI = PARB(I,3)
          ACI = PARC(I,1)
          BCI = PARC(I,2)
          CCI = PARC(I,3)
          DO 250 J = 1,20
              TJ = MIDT(J)
              TTJ = TJ * A + B
              WJ = WEIGHT(J)
              IF (WJ .EQ. 0.0) GOTO 250
C
C      PARTIAL DERIVATIVE OF PB WITH RESPECT TO B
      ANS=(-EXP(BBI*ABI*D+TJ*ABI*D*A+ABI*D*B)*(CBI-1.)*ABI*D)/(EXP
     + (BBI*ABI*D)+EXP(TJ*ABI*D*A+ABI*D*B))**2
C
C      COMPUTE PROBS
      PB = P3PL(ABI,BBI,CBI,TTJ)
      PC = P3PL(ACI,BCI,CCI,TJ)
C
C      COMPUTE DIRIVITIVE OF TERM
      TERM = 2. * WJ * ANS * (PB - PC)
C
      PDB = PDB + TERM
 250          CONTINUE
 300      CONTINUE
C
      RETURN
      END
C
C ************************** MLE EQUATING **************************
C
      SUBROUTINE EQUMLE(PARB,PARC,IROW,NITEMS,THETAB,THETAC,
     * NSUBSB,NSUBSC,NPAR,A,B,IOUT,IER)
C
      REAL PARB(IROW,3),PARC(IROW,3),THETAB(NSUBSB),THETAC(NSUBSC),
     * COVB(3,3,200),COV1(3,3),COV2(3,3),COVC(3,3,200),
     * X(2),
     * H(3),G(2),W(6),A(7),B(7),SPARB(200,3),SPARC(200,3)
      INTEGER SNITEM,SNPAR
C
      EXTERNAL MLEFUN
      COMMON COVB,COVC,SNITEM,SPARB,SPARC,SNPAR
C
C      INITIALIZE ERROR VAR
      IER = 0
C
C      INITIALIZE STARTING VALUES OF EQUATING CONSTANTS
```

```
      IF (A(2) .EQ. 0.0 .AND. B(2) .EQ. 0.0)
     * CALL BEQUAR(PARB,PARC,IROW,NITEMS,A,B,IOUT,IER)
      X(1) = A(2)
      X(2) = B(2)
      CALL UGETIO(3,0,5)
C
C     TRANSFER STUFF TO COMMON ARRAYS
      DO 441 I = 1,NITEMS
          DO 411 J = 1,3
              SPARB(I,J) = PARB(I,J)
              SPARC(I,J) = PARC(I,J)
  411         CONTINUE
  441     CONTINUE
      SNITEM = NITEMS
      SNPAR = NPAR
C
C     COMPUTE COVARIANCE MATRICES
      DO 99 ITM = 1,NITEMS
          AB = PARB(ITM,1)
          BB = PARB(ITM,2)
          CB = PARB(ITM,3)
          AC = PARC(ITM,1)
          BC = PARC(ITM,2)
          CC = PARC(ITM,3)
          CALL COV3PL(THETAB,NSUBSB,AB,BB,CB,NPAR,3,COV1,IOUT,IER)
          CALL COV3PL(THETAC,NSUBSC,AC,BC,CC,NPAR,3,COV2,IOUT,IER)
          DO 71 I = 1,NPAR
              DO 69 J = 1,NPAR
                  COVB(I,J,ITM) = COV1(I,J)
                  COVC(I,J,ITM) = COV2(I,J)
   69             CONTINUE
   71         CONTINUE
   99     CONTINUE
C
C     PERFORM MINIMIZATION
      NSIG = 3
      N = 2
      MAXFN = 500
      IOPT = 2
      CALL ZXMIN(MLEFUN,N,NSIG,MAXFN,IOPT,X,H,G,F,W,IER)
C
      A(7) = X(1)
      B(7) = X(2)
C
      RETURN
      END
C
      SUBROUTINE MLEFUN(N,X,F)
      REAL COVB(3,3,200),COVC(3,3,200),SPARB(200,3),
     * SPARC(200,3),COV(3,3),WK(3),SCOV(3,3),
     * V(3),X(2)
      INTEGER SNITEM,SNPAR
```

```
      COMMON COVB,COVC,SNITEM,SPARB,SPARC,SNPAR
C
C     INITIALIZE VALUES
      A = X(1)
      B = X(2)
      F = 1.
      KK = 0
C
      DO 500 ITM = 1,SNITEM
C
C     TRANSFORM COVARIANCE MATRIX FOR COMPARISON GROUP
      DO 9 I = 1,SNPAR
          DO 7 J = 1,SNPAR
              COV(I,J) = COVC(I,J,ITM)
    7         CONTINUE
    9     CONTINUE
      COV(1,1) = COV(1,1) / (A**2.)
      COV(2,2) = COV(2,2) * (A**2.)
      IF (SNPAR .EQ. 2) GOTO 39
      COV(1,3) = COV(1,3) / A
      COV(3,1) = COV(1,3)
      COV(2,3) = COV(2,3) * A
      COV(3,2) = COV(2,3)
   39 CONTINUE
C
C     COMPUTE SUM OF COVARIANCE MATRICES
      DO 45 I = 1,SNPAR
          DO 43 J = 1,SNPAR
              SCOV(I,J) = COV(I,J) + COVB(I,J,ITM)
   43         CONTINUE
   45     CONTINUE
C
C     TRANSFORM ITEM PARAMETERS FOR COMPARISON GROUP
      APAR = SPARC(ITM,1) / A
      BPAR = (SPARC(ITM,2) * A) + B
C
C     COMPUTE VECTOR OF PARAMETER DIFFERENCES
      V(1) = SPARB(ITM,1) - APAR
      V(2) = SPARB(ITM,2) - BPAR
      V(3) = SPARB(ITM,3) - SPARC(ITM,3)
C
C     COMPUTE MULTIVARIATE DENSITY VALUE (PROB)
      CALL NORDEN(SCOV,V,SNPAR,PROB,CS)
C
C     INCREMENT CURRENT FUNCTION VALUE
      F = F * PROB
C
C     CHECK AND CORRECT FOR UNDERFLOW
   70 IF(ABS(F) .GT. 1.) GOTO 100
      KK = KK + 1
      F = F * 1024.
      GOTO 70
```

```
 100 CONTINUE
C
 500 CONTINUE
C
C     COMPUTE LOG OF FUNCTION VALUE AND MULTIPLY BY -1
      F = - ALOG(F) + 10. * FLOAT(KK) * ALOG(2.)
C
      RETURN
      END
C
      SUBROUTINE  NORDEN(COV,CEN,P,DEN,CS)
      INTEGER P
      REAL COV(3,3),CEN(3),WK(3),INVCOV(3,3),B(3),TEMP(3,3),
     * TEMP2(3)
C
C     INVERT COVARIANCE MATRIX
      DO 11 I = 1,P
          DO 10 J = 1,P
              TEMP(I,J) = COV(I,J)
   10         CONTINUE
   11     CONTINUE
      CALL LINV1F(TEMP,P,3,INVCOV,0,WK,IER)
C
C     COMPUTE CHI-SQUARE
      DO 79 I = 1,P
          TEMP2(I) = 0.
          DO 77 J = 1,P
              TEMP2(I) = TEMP2(I) + CEN(J) * INVCOV(J,I)
   77         CONTINUE
   79     CONTINUE
      CS = 0.
      DO 48 K = 1,P
          CS = CS + TEMP2(K) * CEN(K)
   48     CONTINUE
C
C     COMPUTE DETERMINANT OF COVARIANCE MATRIX
      D1 = 5.
      CALL LINV3F(COV,B,4,P,3,D1,D2,WK,IER)
      DET = D1 * (2.**D2)
C
C     COMPUTE P-VARIATE NORMAL DENSITY
      PIE = 3.1415927
      DEN = ((2.*PIE)**(-P/2.)) * (DET**(-.5)) * EXP(-CS/2.)
C
      RETURN
      END
c
c ********************** UTILITY SUBROUTINES ***************************
c
P3PL  -  Function
```

This routine computes probabilities according to the 3-parameter logistic model for given values of a,b,c and theta.

REAL FUNCTION P3PL(A,B,C,THETA)

A: a-parameter

B: b-parameter

C: c-parameter

THETA: Person parameter

Function listing:

```
      REAL FUNCTION  P3PL(A,B,C,THETA)
      XNUM = 1. - C
      DENOM = 1. + EXP(-1.702 * A * (THETA - B))
      P3PL = C + (XNUM/DENOM)
      RETURN
      END
```

ESPNT - SUBROUTINE

THIS SUBROUTINE CREATES A VECTOR OF EQUALY SPACED POINTS

ESPNT(X,NPONTS,XMIN,XMAX)

X: OUTPUT VECTOR OF LENGTH NPONTS WHICH CONTAINS THE EQUALY SPACED POINTS.

NPONTS: NUMBER OF POINTS WHICH X WILL CONTAIN. NPONTS SHOULD BE GREATER THAN OR EQUAL TO 2.

XMIN: MINIMUM VALUE OF X VECTOR. WILL BE PLACED IN FIRST ELEMENT OF VECTOR X.

XMAX: MAXIMUM VALUE OF X VECTOR. WILL BE PLACED IN LAST ELEMENT OF VECTOR X.

Subroutine listing:

```
      SUBROUTINE ESPNT(X,NPONTS,XMIN,XMAX)
      REAL X(NPONTS)
C
C
      X(1) = XMIN
      XPONTS = NPONTS
      XNINT = XPONTS - 1.
      XINCRE = (XMAX-XMIN)/XNINT
      DO 10 I = 2,NPONTS
          X(I) = X(I-1) + XINCRE
   10     CONTINUE
      RETURN
      END
```

COV3PL - SUBROUTINE

THIS ROUTINE COMPUTES THE ITEM COVARIANCE MATRIX FOR THE TWO OR THREE PARAMETER LOGISTIC MODELS. FORMULAS ARE TAKEN FROM LORD (1980) P.191.

COV3PL (THETAS,NSUBS,A,B,C,NPAR,IR,COV,IOUT,IER)

THETAS: VECTOR OF LENGTH NSUBS CONTAINING THE THETA PARAMETERS FOR THOSE ANSWERING THE ITEM. IF A SUBJECT DID NOT ANSWER THE ITEM, THE THETA FOR THAT PERSON SHOULD BE SET TO 999.

NSUBS: LENGTH OF VECTOR THETAS, INCLUDING 999'S. NUMBER OF SUBJECTS.

A: A-PARAMETER FOR THE THREE PARAMETER LOGISTIC MODEL.

B: B-PARAMETER FOR THE THREE PARAMETER LOGISTIC MODEL.

C: C-PARAMETER FOR THE THREE PARAMETER LOGISTIC MODEL. IF THE TWO PARAMETER MODEL IS DESIRED, C SHOULD BE SET EQUAL TO 0.

NPAR: NUMBER OF ESTIMATED PARAMETERS IN MODEL. NPAR=2 FOR THE TWO PARAMETER MODEL, OR WHEN C IS KNOWN, NPAR=3 FOR THE THREE PARAMETER MODEL.

IR: ROW DIMENSION OF MATRIX COV EXACTLY AS SPECIFIED IN THE

```
         CALLING PROGRAM.

COV:     OUTPUT, NPAR BY NPAR MATRIX CONTAINING THE COVARIANCE
         MATRIX.

IOUT:    TAPE NUMBER OF OUTPUT FILE.

IER:     IF IER > 0, AN ERROR OCCURED IN AN IMSL SUBROUTINE AND
         RESULTS IN COV ARE NOT THE ACTUAL ESTIMATES.
```

Subroutine listing:

```
      SUBROUTINE  COV3PL(THETAS,NSUBS,A,B,C,NPAR,IR,COV,IOUT,IER)
      REAL THETAS(NSUBS),COV(IR,3),
     * INF(3,3),WK(3),TCOV(3,3)
      CALL UGETIO(3,0,IOUT)
C
C      INITIALIZE VALUES TO ZERO
      DO 30 J = 1,NPAR
          DO 25 I = 1,NPAR
              INF(I,J) = 0.0
   25         CONTINUE
   30     CONTINUE
C
C      COMPUTE ADDITIVE TERMS IN COVARIANCE MATRIX
      DO 100 J = 1,NSUBS
          THET = THETAS(J)
          IF (THET .EQ. 999.0) GOTO 100
          PROB = P3PL(A,B,C,THET)
          QDP = (1.-PROB)/PROB
C
          INF(1,1) = INF(1,1) + (((THET-B)**2.) * ((PROB-C)**2.) *
     *     QDP)
          INF(2,2) = INF(2,2) + (((PROB-C)**2.) * QDP)
          INF(2,1) = INF(2,1) + ((THET-B) * ((PROB-C)**2.) * QDP)
C
          IF (NPAR .EQ. 2) GOTO 100
          INF(3,1) = INF(3,1) + ((THET-B) * (PROB-C) * QDP)
          INF(3,2) = INF(3,2) + ((PROB-C) * QDP)
          INF(3,3) = INF(3,3) + QDP
  100     CONTINUE
C
C      MULTIPLY TERMS BY RESPECTIVE CONSTANTS
      D = 1.702
      CTERM = 1./((1.-C)**2.)
      INF(1,1) = (D**2.) * CTERM * INF(1,1)
      INF(2,1) = (D**2.) * A * CTERM * INF(2,1) * (-1.)
      INF(1,2) = INF(2,1)
      INF(2,2) = (D**2.) * (A**2.) * CTERM * INF(2,2)
C
```

```
      IF (NPAR .EQ. 2) GOTO 200
      INF(3,1) = D * CTERM * INF(3,1)
      INF(1,3) = INF(3,1)
      INF(3,2) = (-1.) * D * A * CTERM * INF(3,2)
      INF(2,3) = INF(3,2)
      INF(3,3) = CTERM * INF(3,3)
  200 CONTINUE
C
C     FIND INVERSE OF INFORMATION MATRIX
      IER = 0
      CALL LINV1F(INF,NPAR,3,TCOV,0,WK,IER)
C
C     COPY RESULTS TO MATRIX: COV
      DO 305 J = 1,NPAR
          DO 300 I = 1,NPAR
              COV(I,J) = TCOV(I,J)
  300         CONTINUE
  305     CONTINUE
C
      RETURN
      END
```

REFERENCES

Bianchini, J. C., & Loret, P. G. Anchor Test Study Final Report. Project report and volumes 1 through 30, and Anchor test study supplement. Volumes 31 through 33. 1974. (Eric Document Reproduction Service Numbers ED 092 601 through ED 092 634).

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison Wesley, 1968.

Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., & Balow, I. H. Metropolitan achievement tests, Form F. New York: Harcourt, Brace, & Jovanovich, 1970.

Haebara, T. Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research, 1980, 22, 144-149.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. An investigation of item bias in a test of reading comprehension (Technical Report No. 163). Urbana, Ill.: Center for the Study of Reading, University of Illinois, 1980.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J An investigation of item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 4.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison Wesley, 1968.

Mckinley, R. L., & Reckase, M. D. A comparison of procedures for constructing large item pools (Research Report 81-3). Tailored Testing Research Laboratory, University of Missouri, 1981.

Segall, D. O. A technique for transforming to a common metric in IRT using vectors of estimated item parameter differences. Manuscript in preparation. 1983.

Segall, D. O., & Levine, M. V. A technique for transforming to a common metric in IRT using a weighted least squares approach. Manuscript in preparation. 1983.

Stocking, M. L., & Lord, F. M. Developing a common metric in item response theory. Research Memorandum 82-25. Princeton, N.J.: Educational Testing Service, 1982.

Wood, R. L., & Lord, F. M. A user's guide to LOGIST. Research Memorandum 76-4. Princeton, N.J.: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST-A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, N.J.: Educational Testing Service, 1976.

VITA

Address

Office: Department of Psychology
University of Illinois
603 East Daniel
Champaign, Illinois 61820

Home: 604 East While, Apt. 23
Champaign, Illinois 61820

Personal Record

Date of Birth: June 2, 1957
Birthplace: Los Angeles, California

Education

1979	B.A.	University of California. Major Area: Psychology. Graduated with High Honors
1981	A.M.	University of Illinois, Urbana-Champaign. Major Area: Quantitative Psychology
1983	Ph.D.	University of Illinois, Urbana-Champaign. Major Area: Quantitative Psychology

Experience

1978-1979	Research Assistant: University of California, Department of Psychology, Dr. Howard Friedman
1979-1980	Teaching Assistant: University of Illinois, Department of Psychology, Dr. Faruk Saad and Dr. Charles Hulin
1980-1981	Research Data Analyst: Survey Research Laboratory, University of Illinois
1981-date	Research Assistant: University of Illinois, Department of Educational Psychology, Dr. Michael Levine

Paper Presented

Levine, M. V. & Segall, D. O. Identifying different response functions. Presented at the American Psychological Association meeting, Washington D. C., August, 1982.

Distribution List

Dr. Terry Ackerman
American College Testing Programs
P.O. Box 168
Iowa City, IA 52243

Dr. Robert Ahlers
Code N711
Human Factors Laboratory
Naval Training Systems Center
Orlando, FL 32813

Dr. James Algina
1403 Norman Hall
University of Florida
Gainesville, FL 32605

Dr. Erling B. Andersen
Department of Statistics
Studiestraede 6
1455 Copenhagen
DENMARK

Dr. Eva L. Baker
UCLA Center for the Study
of Evaluation
145 Moore Hall
University of California
Los Angeles, CA 90024

Dr. Isaac Bejar
Mail Stop: 10-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Menucha Birenbaum
School of Education
Tel Aviv University
Ramat Aviv 69978
ISRAEL

Dr. Arthur S. Blaiwes
Code N712
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Bruce Bloxom
Defense Manpower Data Center
550 Camino El Estero,
Suite 200
Monterey, CA 93943-3231

Dr. R. Darrell Bock
University of Chicago
NORC
6030 South Ellis
Chicago, IL 60637

Cdt. Arnold Bohrer
Sectie Psychologisch Onderzoek
Rekruterings-En Selectiecentrum
Kwartier Koningen Astrid
Bruijnstraat
1120 Brussels, BELGIUM

Dr. Robert Breaux
Code 7B
Naval Training Systems Center
Orlando, FL 32813-7100

Dr. Robert Brennan
American College Testing
Programs
P. O. Box 168
Iowa City, IA 52243

Dr. James Carlson
American College Testing
Program
P.O. Box 168
Iowa City, IA 52243

Dr. John B. Carroll
409 Elliott Rd., North
Chapel Hill, NC 27514

Dr. Robert M. Carroll
Chief of Naval Operations
OP-01B2
Washington, DC 20350

Distribution List

Dr. Raymond E. Christal
UES LAMP Science Advisor
AFHRL/MOEL
Brooks AFB, TX 78235

Dr. Norman Cliff
Department of Psychology
Univ. of So. California
Los Angeles, CA 90089-1061

Director,
Manpower Support and
Readiness Program
Center for Naval Analysis
2000 North Beauregard Street
Alexandria, VA 22311

Dr. Stanley Collyer
Office of Naval Technology
Code 222
800 N. Quincy Street
Arlington, VA 22217-5000

Dr. Hans F. Crombag
Faculty of Law
University of Limburg
P.O. Box 616
Maastricht
The NETHERLANDS 6200 MD

Dr. Timothy Davey
Educational Testing Service
Princeton, NJ 08541

Dr. C. M. Dayton
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Ralph J. DeAyala
Measurement, Statistics,
and Evaluation
Benjamin Bldg., Rm. 4112
University of Maryland
College Park, MD 20742

Dr. Dattprasad Divgi
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. Hei-Ki Dong
Bell Communications Research
6 Corporate Place
PYA-1K226
Piscataway, NJ 08854

Dr. Fritz Drasgow
University of Illinois
Department of Psychology
603 E. Daniel St.
Champaign, IL 61820

Defense Technical
Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

Dr. Stephen Dunbar
224B Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. James A. Earles
Air Force Human Resources Lab
Brooks AFB, TX 78235

Dr. Kent Eaton
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. John M. Eddins
University of Illinois
252 Engineering Research
Laboratory
103 South Mathews Street
Urbana, IL 61801

Distribution List

Dr. Susan Embretson
University of Kansas
Psychology Department
426 Fraser
Lawrence, KS 66045

Dr. George Englehard, Jr.
Division of Educational Studies
Emory University
210 Fishburne Bldg.
Atlanta, GA 30322

Dr. Benjamin A. Fairbank
Performance Metrics, Inc.
5825 Callaghan
Suite 225
San Antonio, TX 78228

Dr. P-A. Federico
Code 51
NPRDC
San Diego, CA 92152-6800

Dr. Leonard Feldt
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Richard L. Ferguson
American College Testing
P.O. Box 168
Iowa City, IA 52243

Dr. Gerhard Fischer
Liebiggasse 5/3
A 1010 Vienna
AUSTRIA

Dr. Myron Fischl
U.S. Army Headquarters
DAPE-MRR
The Pentagon
Washington, DC 20310-0300

Prof. Donald Fitzgerald
University of New England
Department of Psychology
Armidale, New South Wales 2351
AUSTRALIA

Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Alfred R. Fregly
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Robert D. Gibbons
Illinois State Psychiatric Inst.
Rm 529W
1601 W. Taylor Street
Chicago, IL 60612

Dr. Janice Gifford
University of Massachusetts
School of Education
Amherst, MA 01003

Dr. Robert Glaser
Learning Research
& Development Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh, PA 15260

Dr. Bert Green
Johns Hopkins University
Department of Psychology
Charles & 34th Street
Baltimore, MD 21218

DORNIER GMBH
P.O. Box 1420
D-7990 Friedrichshafen 1
WEST GERMANY

Distribution List

Dr. Ronald K. Hambleton
University of Massachusetts
Laboratory of Psychometric
and Evaluative Research
Hills South, Room 152
Amherst, MA 01003

Dr. Delwyn Harnisch
University of Illinois
51 Gerty Drive
Champaign, IL 61820

Dr. Grant Henning
Senior Research Scientist
Division of Measurement
Research and Services
Educational Testing Service
Princeton, NJ 08541

Ms. Rebecca Hetter
Navy Personnel R&D Center
Code 63
San Diego, CA 92152-6800

Dr. Paul W. Holland
Educational Testing Service, 21-T
Rosedale Road
Princeton, NJ 08541

Prof. Lutz F. Hornke
Institut fur Psychologie
RWTH Aachen
Jaegerstrasse 17/19
D-5100 Aachen
WEST GERMANY

Dr. Paul Horst
677 G Street, #184
Chula Vista, CA 92010

Mr. Dick Hoshaw
OP-135
Arlington Annex
Room 2834
Washington, DC 20350

Dr. Lloyd Humphreys
University of Illinois
Department of Psychology
603 East Daniel Street
Champaign, IL 61820

Dr. Steven Hunka
3-104 Educ. N.
University of Alberta
Edmonton, Alberta
CANADA T6G 2G5

Dr. Huynh Huynh
College of Education
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Jannarone
Elec. and Computer Eng. Dept.
University of South Carolina
Columbia, SC 29208

Dr. Douglas H. Jones
Thatcher Jones Associates
P.O. Box 6640
10 Trafalgar Court
Lawrenceville, NJ 08648

Dr. Milton S. Katz
European Science Coordination
Office
U.S. Army Research Institute
Box 65
FPO New York 09510-1500

Prof. John A. Keats
Department of Psychology
University of Newcastle
N.S.W. 2308
AUSTRALIA

Dr. G. Gage Kingsbury
Portland Public Schools
Research and Evaluation Department
501 North Dixon Street
P. O. Box 3107
Portland, OR 97209-3107

Distribution List

Dr. William Koch
Box 7246, Meas. and Eval. Ctr.
University of Texas-Austin
Austin, TX 78703

Dr. James Kraatz
Computer-based Education
Research Laboratory
University of Illinois
Urbana, IL 61801

Dr. Leonard Kroeker
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Dr. Jerry Lehnus
Defense Manpower Data Center
Suite 400
1600 Wilson Blvd
Rosslyn, VA 22209

Dr. Thomas Leonard
University of Wisconsin
Department of Statistics
1210 West Dayton Street
Madison, WI 53705

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Charles Lewis
Educational Testing Service
Princeton, NJ 08541-0001

Dr. Robert L. Linn
Campus Box 249
University of Colorado
Boulder, CO 80309-0249

Dr. Robert Lockman
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. Frederic M. Lord
Educational Testing Service
Princeton, NJ 08541

Dr. George B. Macready
Department of Measurement
Statistics & Evaluation
College of Education
University of Maryland
College Park, MD 20742

Dr. Gary Marco
Stop 31-E
Educational Testing Service
Princeton, NJ 08451

Dr. James R. McBride
The Psychological Corporation
1250 Sixth Avenue
San Diego, CA 92101

Dr. Clarence C. McCormick
HQ, USMEPCOM/MEPCT
2500 Green Bay Road
North Chicago, IL 60064

Dr. Robert McKinley
Educational Testing Service
16-T
Princeton, NJ 08541

Dr. James McMichael
Technical Director
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Barbara Means
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025

Distribution List

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Dr. William Montague
NPRDC Code 13
San Diego, CA 92152-6800

Ms. Kathleen Moreno
Navy Personnel R&D Center
Code 62
San Diego, CA 92152-6800

Headquarters Marine Corps
Code MPI-20
Washington, DC 20380

Dr. W. Alan Nicewander
University of Oklahoma
Department of Psychology
Norman, OK 73071

Deputy Technical Director
NPRDC Code 01A
San Diego, CA 92152-6800

Director, Training Laboratory,
NPRDC (Code 05)
San Diego, CA 92152-6800

Director, Manpower and Personnel
Laboratory,
NPRDC (Code 06)
San Diego, CA 92152-6800

Director, Human Factors
& Organizational Systems Lab,
NPRDC (Code 07)
San Diego, CA 92152-6800

Library, NPRDC
Code P201L
San Diego, CA 92152-6800

Commanding Officer,
Naval Research Laboratory
Code 2627
Washington, DC 20390

Dr. Harold F. O'Neil, Jr.
School of Education - WPH 801
Department of Educational
Psychology & Technology
University of Southern California
Los Angeles, CA 90089-0031

Dr. James B. Olsen
WICAT Systems
1875 South State Street
Orem, UT 84058

Office of Naval Research,
Code 1142CS
800 N. Quincy Street
Arlington, VA 22217-5000

Office of Naval Research,
Code 125
800 N. Quincy Street
Arlington, VA 22217-5000

Assistant for MPT Research,
Development and Studies
OP 01B7
Washington, DC 20370

Dr. Judith Orasanu
Basic Research Office
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard St.
Alexandria, VA 22311

Distribution List

Dr. Randolph Park
Army Research Institute
5001 Eisenhower Blvd.
Alexandria, VA 22333

Wayne M. Patience
American Council on Education
GED Testing Service, Suite 20
One Dupont Circle, NW
Washington, DC 20036

Dr. James Paulson
Department of Psychology
Portland State University
P.O. Box 751
Portland, OR 97207

Dept. of Administrative Sciences
Code 54
Naval Postgraduate School
Monterey, CA 93943-5026

Department of Operations Research,
Naval Postgraduate School
Monterey, CA 93940

Dr. Mark D. Reckase
ACT
P. O. Box 168
Iowa City, IA 52243

Dr. Malcolm Ree
AFHRL/MOA
Brooks AFB, TX 78235

Dr. Barry Riegelhaupt
HumRRO
1100 South Washington Street
Alexandria, VA 22314

Dr. Carl Ross
CNET-PDCD
Building 90
Great Lakes NTC, IL 60088

Dr. J. Ryan
Department of Education
University of South Carolina
Columbia, SC 29208

Dr. Fumiko Samejima
Department of Psychology
University of Tennessee
310B Austin Peay Bldg.
Knoxville, TN 37916-0900

Mr. Drew Sands
NPRDC Code 62
San Diego, CA 92152-6800

Lowell Schoer
Psychological & Quantitative
Foundations
College of Education
University of Iowa
Iowa City, IA 52242

Dr. Mary Schratz
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Dan Segall
Navy Personnel R&D Center
San Diego, CA 92152

Dr. W. Steve Sellman
OASD(MRA&L)
2B269 The Pentagon
Washington, DC 20301

Dr. Kazuo Shigemasu
7-9-24 Kugenuma-Kaigan
Fujisawa 251
JAPAN

Dr. William Sims
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Distribution List

Dr. H. Wallace Sinaiko
Manpower Research
and Advisory Services
Smithsonian Institution
801 North Pitt Street, Suite 120
Alexandria, VA 22314-1713

Dr. Richard E. Snow
School of Education
Stanford University
Stanford, CA 94305

Dr. Richard C. Sorensen
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Paul Speckman
University of Missouri
Department of Statistics
Columbia, MO 65201

Dr. Judy Spray
ACT
P.O. Box 168
Iowa City, IA 52243

Dr. Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Dr. Peter Stoloff
Center for Naval Analysis
4401 Ford Avenue
P.O. Box 16268
Alexandria, VA 22302-0268

Dr. William Stout
University of Illinois
Department of Statistics
101 Illini Hall
725 South Wright St.
Champaign, IL 61820

Dr. Hariharan Swaminathan
Laboratory of Psychometric and
Evaluation Research
School of Education
University of Massachusetts
Amherst, MA 01003

Mr. Brad Sympson
Navy Personnel R&D Center
Code-62
San Diego, CA 92152-6800

Dr. John Tangney
AFOSR/NL, Bldg. 410
Bolling AFB, DC 20332-6448

Dr. Kikumi Tatsuoka
CERL
252 Engineering Research
Laboratory
103 S. Mathews Avenue
Urbana, IL 61801

Dr. Maurice Tatsuoka
220 Education Bldg
1310 S. Sixth St.
Champaign, IL 61820

Dr. David Thissen
Department of Psychology
University of Kansas
Lawrence, KS 66044

Mr. Gary Thomasson
University of Illinois
Educational Psychology
Champaign, IL 61820

Dr. Robert Tsutakawa
University of Missouri
Department of Statistics
222 Math. Sciences Bldg.
Columbia, MO 65211

Distribution List

Dr. Ledyard Tucker
University of Illinois
Department of Psychology
603 E. Daniel Street
Champaign, IL 61820

Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E. Street, NW
Washington, DC 20415

Dr. David Vale
Assessment Systems Corp.
2233 University Avenue
Suite 440
St. Paul, MN 55114

Dr. Frank L. Vicino
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Howard Wainer
Educational Testing Service
Princeton, NJ 08541

Dr. Ming-Mei Wang
Lindquist Center
for Measurement
University of Iowa
Iowa City, IA 52242

Dr. Thomas A. Warm
Coast Guard Institute
P. O. Substation 18
Oklahoma City, OK 73169

Dr. Brian Waters
HumRRO
12908 Argyle Circle
Alexandria, VA 22314

Dr. David J. Weiss
N660 Elliott Hall
University of Minnesota
75 E. River Road
Minneapolis, MN 55455-0344

Dr. Ronald A. Weitzman
Box 146
Carmel, CA 93921

Major John Welsh
AFHRL/MOAN
Brooks AFB, TX 78223

Dr. Douglas Wetzel
Code 51
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. Rand R. Wilcox
University of Southern
California
Department of Psychology
Los Angeles, CA 90089-1061

German Military Representative
ATTN: Wolfgang Wildgrube
Streitkraefteamt
D-5300 Bonn 2
4000 Brandywine Street, NW
Washington, DC 20016

Dr. Bruce Williams
Department of Educational
Psychology
University of Illinois
Urbana, IL 61801

Dr. Hilda Wing
NRC MH-176
2101 Constitution Ave.
Washington, DC 20418

Distribution List

Dr. Martin F. Wiskoff
Defense Manpower Data Center
550 Camino El Estero
Suite 200
Monterey, CA 93943-3231

Mr. John H. Wolfe
Navy Personnel R&D Center
San Diego, CA 92152-6800

Dr. George Wong
Biostatistics Laboratory
Memorial Sloan-Kettering
Cancer Center
1275 York Avenue
New York, NY 10021

Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
Code 51
San Diego, CA 92152-6800

Dr. Kentaro Yamamoto
03-T
Educational Testing Service
Rosedale Road
Princeton, NJ 08541

Dr. Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

Dr. Joseph L. Young
National Science Foundation
Room 320
1800 G Street, N.W.
Washington, DC 20550

Mr. Anthony R. Zara
National Council of State
Boards of Nursing, Inc.
625 North Michigan Avenue
Suite 1544
Chicago, IL 60611